Smooth regression quantile estimation

Thesis

The Open University

Smooth Regression Quantile Estimation

Keming Yu, B.Sc., M.Phil.

A thesis submitted for the degree of Doctor of Philosophy in the Faculty of Mathematics and Computing of The Open University

Department of Statistics, The Open university

1996

Preface

This thesis is based on the findings of research carried out in the Department of Statistics, The Open University, Walton Hall, Milton Keynes. The contents of this thesis are original, except where specific references are made to other work.

No part of this thesis has been submitted for a degree or other qualifications to any other University.

Acknowledgements

I am glad to have this opportunity to express my deepest gratitude to my supervisor, Dr. M.C. Jones. Throughout the past three years there has rarely been a day that he was not available to help or advise me. Particularly, I have greatly benefited from weekly meetings with him in the first two years. He guided my steps in subject with clarity, patience and continuous encouragement.

I am sincerely thankful to Prof. David Hand whose kind help and advice has lead to improvement and clarification of the results. I am much indebted to Dr Emmanuel Aziz for corrections of English and comments on the first draft of this thesis.

I thank Dr T. J. Cole (MRC Dunn Nutrition of Cambridge), Dr P. Royston and Dr E. Wright (Royal Postgraduate Medical School of London) for providing the data sets , Dr M. P. Wand (University of New South Wales, Australian) for S code to implement Ruppert–Sheather–Wand bandwidth selection method and Ms H. Pan (London University) for supplying their manual copy of GROSTAT .

I am thankful to all the staffs of statistics department of the Open University for their assistance and advice, and I gratefully acknowledge the Open University support through a research studentship of which the present work is the outcome.

Finally, I thank my wife, Ling Wang, for her support, confidence and love through all those long tiring times and our parents for their constant encouragement which gave me the strength to overcome every difficulty I encountered.

Summary

This thesis deals with some problems of nonparametric smoothing of regression quantiles (conditional quantiles) particularly kernel smoothing using local linear fitting, with some emphasis on medical applications (smoothing reference or centile charts). New methods and algorithms for practical applications are developed, including automatic bandwidth selection, and methods are compared.

The main aspects of smoothing regression quantiles are summarized below:

1. **Local Linear Single-Kernel Smoothing Regression Quantiles Defined by "Check Function"**

   Unlike smoothing the regression mean, smoothing parameter selection and practical computation issues of nonparametric smoothing regression quantiles have received little attention in the statistical literature. This thesis gives a novel idea for bandwidth selection and an algorithm in the field of nonparametric smoothing regression quantiles based on kernel weighted local linear fitting and regression quantile as the minimizer of $E\{\rho_p(Y-a)|X=x\}$ and ρ_p is an appropriate "check" function for $0<p<1$.

2. **Local Linear Double-Kernel Smoothing Regression Quantiles Defined by "Conditional Distribution Function"**

   By extending the idea of double-kernel smoothing conditional distribution density, this thesis develops a new smoothing approach with kernel weighted local linear fitting for conditional quantiles and gives

some rule-of-thumb for the second bandwidth selection. The advantages of the method are that in practice the centile curves do not cross and it has smaller mean square error than single-kernel fitting.

3. **Comparison of Smoothing Conditional Distribution Function and Its Quantiles Fitted with Local Constant and Local Linear Approaches**

It is known that there are some differences in smoothing regression mean by local constant and local linear fitting, the question now is how different are the smoothing regression quantiles using the two fittings? This thesis makes a good discussion of this comparison.

4. **Local Linear Likelihood-Based Quantile Smoothing**

By developing theory of multi-parameter likelihood-based model, a general structure of polynomial fitting quantiles, variable transformation, and multi-parameter likelihood-based models is discussed. Particularly, a kernel-version of Cole's method and other details are proposed.

5. $k - NN$ **and Local Linear Fitting**

Like smoothing regression mean and density, the $k - NN$ method is an important one but there are disadvantages in smoothing regression quantiles. However, combining it with local linear fitting, its performance is improved, resulting almost in an equivalent version of local linear single-kernel minimizing "check function".

6. Local Polynomial Fit with Penalized Least-Squares Smoothing Regression Function

This thesis propose the idea of estimating the regression function at a particular point by local polynomial fit with roughness penalty. We investigate the structure, algorithm and asymptotic properties. These results show the method has the advantages of usual spline smoothing with $(2k-1)$ order spline penalized by kth derivative.

Outline of the thesis

In this thesis, attention will be mainly focused on the local linear kernel regression quantile estimation. Different estimators within this class have been proposed, developed asymptotically and applied to real applications. I include algorithm-design and selection of smoothing parameters.

Chapter 2 studies two estimators, first a single-kernel estimator based on "check function" and a bandwidth selection rule is proposed based on the asymptotic MSE of this estimator. Second a recursive double-kernel estimator which extends Fan *et al*'s (1996) density estimator, and two algorithms are given for bandwidth selection.

In Chapter 3, a comparison is carried out of local constant fitting and local linear fitting using MSEs of the estimates as a criterion.

Chapter 4 gives a theoretical summary and a simulation study of local linear kernel estimation of conditional distribution function. This has a special interest in itself as well as being related to regression quantiles.

In Chapter 5, a kernel-version method of LMS (Cole and Green, 1992) is considered. The method proposed, which is still a semi-parametric one, is based on a general idea of local linear kernel approach of log-likelihood model.

Chapter 6 proposes a two-step method of smoothing regression quantiles called BPK. The method considered is based on the idea of combining $k - NN$ method with Healy's *et al* (1988) partition rule, and correlated regression model are involved.

In Chapter 7, methods of regression quantile estimation are compared for different underlying models and design densities in a simulation study. The ISE criterion of interior and boundary points is used as a basis for these comparisons. Three methods are recommended for quantile regression in practice, and they are double kernel method, LMS method and Box partition kernel method (BPK).

In Chapter 8, attention is turned to a novel idea of local polynomial roughness penalty regression model, where a purely theoretical framework is considered.

Nomenclature

(X, Y) = bivariate random variables.

$F(x, y)$ = bivariate distribution of (X, Y).

$F(y|x)$ = conditional distribution of Y given $X = x$.

$f(y|x)$ = conditional density of $F(y|x)$.

$g(x)$ = marginal density of X.

$\Phi(.)$ = standard normal distribution.

$\phi(.)$ = standard normal density.

K or W = kernel.

h or b, also h_2 = bandwidth.

$q_p(x)$ = The conditional p-quantile of Y given $X = x$, or regression quantile.

$f(q_p(x)|x) = f(y|x)|_{y=q_p(x)}$

$F(q_p(x)|x) = F(y|x)|_{y=q_p(x)}$

$F^{a,b}(y|x) = \frac{\partial^{a+b}}{\partial u^a \partial v^b} F(v|u)|_{u=x,v=y}$

$I_A(z)$ = indicator function of set A.

$\rho_p(z) = pzI(z > 0) - (1 - p)zI(z \leq 0)$ "check function".

i.i.d = independently and identically distributed.

MSE = mean square error.

MISE = integrated mean square error.

Contents

Chapter 1

Introduction

1.1 General Introduction

In nonparametric estimation of regression function, smooth regression mean curve has received so far much attention in the literature. See Eubank (1988), Müller (1988), Härdle (1990), Wahba (1990), Hastie and Tibshirani (1990) and Wand and Jones (1995) for interesting applications, and good introductions to the general subject area. However, extending applications and gaining new insights about the underlying structures can be obtained by considering other aspects of conditional distributions. An important example of the latter, namely, the nonparametric estimation of quantiles of the conditional distribution, or regression quantiles for short, will be discussed extending recent advances in kernel weighted local polynomial fitting regression mean estimation. Particularly, we lay some stress on smoothing reference charts in medicine.

Let $(X_1, Y_1), (X_2, Y_2), ...$ be random vectors which are iid as (X,Y), Y taking

values in R^1. For $0 < p < 1$, let $q_p(x)$ denote the conditional p-quantile of Y given $X = x$. We consider the problem of estimating $q_p(x)$ from the data $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$, the asymptotic properties of the kernel estimators as $n \to \infty$, and moreover, the development of a really good practical implementation of kernel conditional quantile estimation such as bandwidth selection and algorithm is a priority.

1.1.1 Regression Quantiles

The basic definition of conditional p-quantile $q_p(x)$ is based on the conditional distribution function $F(y|x)$ of Y given X evaluated at $X = x$ of a bivariate vector (X, Y) and assuming $F(.|x)$ is increasing and continuous at x. The p-quantile $q_p(x)$ is the solution of the equation

$$F(q_p(x)|x) = p. \tag{1.1}$$

That is $P(Y \leq q_p(x)|X = x) = p$. Further, if the conditional density function $f(y|x)$ of $F(y|x)$ exists and it is integrable, then, for $X = x$, $q_p(x)$ can be defined as the upper limit of the following integral equation:

$$\int_{-\infty}^{q_p(x)} f(y|x)dy = p. \tag{1.2}$$

A characterisation of $q_p(x)$ is obtained as a function θ that minimises

$$E\{\rho_p(y - \theta)|X = x\},$$

where

$$\rho_p(z) = pzI_{[0,\infty)}(z) - (1 - p)zI_{(-\infty,0)}(z) \tag{1.3}$$

and $I_A(z)$ is the usual indicator function. The pth quantile of a population can be shown to minimize

$$\int [p(y - \theta)I_{[\theta,\infty]}(y) + (1 - p)(\theta - y)I_{[-\infty,\theta]}(y)]dF(y) \tag{1.4}$$

The function $\rho_p(z)$ is called "check function" or specific loss function which is viewed as an extension of regression median in econometric literature. An alternative way of writing it is as

$$\rho_p(z) = \{|z| + (2p-1)z\}/2.$$

An alternative definition of $q_p(x)$ is to estimate the unknown function $\theta(x)$ of the regression model

$$Y = \theta(X) + \epsilon \tag{1.5}$$

where X and ϵ are independent and pth quantile of ϵ is 0. Obviously, these forms are equivalent, which is convenient to develop the theory and methods.

1.1.2 Smoothing Regression Quantiles

Generally speaking, quantiles of a variable Y conditional on another variable X, when plotted against X can be a useful descriptive tool. These plots give a quick impression of the functional form of the relation between X and the location, also spread and the shape of the conditional distribution of Y. The resulting quantile plot may be quite noisy, however, smoothing across X may be desired.

In medicine, the regression quantile is called reference centile, and a collection of reference centiles is a reference chart. The charts are widely used as screening tools to identify unusual subjects in the area of preliminary medical diagnosis in the sense that the value of some particular measurement on these individuals lies in one or other tail of the reference distribution. The need for centile curves rather than a simple reference range arises when the measurement is strongly dependent on some covariate, often age, upon which the reference range depends as well. The chosen centiles are usually a symmetric subset of $\{3, 5, 10, 25, 50, 75, 90, 95, 97\}$.

In predicting the response from a given covariate $X = x$, estimates of $F^{-1}(p/2|x)$ and $F^{-1}(1-p/2|x)$ can be used to obtain a $100(1-p)$ percent nonparametric predictive interval. This naturally is compared with approaches based on parametric models which lack the ability to deal with the bias arising from misspecification of the model.

Also, they are useful for assessing departure from model assumptions, especially "heteroscedasticity" or other forms of error heterogeneity in generalized linear models. For example, if data points are identically distributed, the conditional quantiles are parallel for different p. Otherwise they will be non-parallel, and the extent of lack of parallelism provides a test for heteroscedasticity (Koenker and Bassett (1982); Efron (1991); Portnoy (1991)).

In short, various percentage points can give a more complete picture of the underlying structure of the data than a grand summary of the averages of the distributions.

Some specific quantiles are more robust than average summary. In fact, it is known that the sample median as an estimator of the mean is more robust than sample mean in classical statistics. Under certain conditions, smoothing regression mean estimation, even by local polynomial smoothing, $\hat{q}_{1/2}(x)$ is more robust than $\hat{m}(x)$.

Alternatively, and because of this, regression mean estimation can be explored using linear combination of sample quantile regression function. See Cheng (1984), Janssen and Veraverbeke (1987) and Lejeune and Sarda (1988).

Finally, specific quantiles may be of independent interest: see Koenker, Portnoy and Ng (1992) and Hendricks and Koenker (1992). For example, discussing elec-

tricity demand (by household) over time in terms of weather characteristics, the low quantile curve corresponds to background use, while the high demand curves reflect the use during active periods of the day (particularly air conditioning).

1.2 Review of the Literature

Many researchers in different areas of application contributed to the study of regression smoothing problems. These studies are generally classified into two wide groups (i) general interest and (ii) medical interest.

1.2.1 Smoothing Conditional Quantiles

Conditional regression quantile functions first were discussed using artificial salary data by Hogg (1975) who regarded the (100p)th percentile of the Y distribution as a regression line $\alpha+\beta x$ of X and called them percentile regression lines. This is actually a parametric estimation approach of conditional quantiles and generalizes the median regression idea of Mood and Brown (Mood, 1950, pp. 406-10).

Koenker and Bassett (1978) generalized the concept and defined pth regression quantile as the minimizing solution of

$$min_{b\in R^K}\Big[\sum_{t\in\{t:y_t\geq x_t b\}} p|y_t - x_t b| + \sum_{t\in\{t:y_t<x_t b\}} (1-p)|y_t - x_t b|\Big] \qquad (1.6)$$

in terms of check function, where $\{x_t : t = 1, ..., T\}$ is a sequence of (row) K-vectors of a known design matrix and $\{y_t : t = 1, ..., T\}$ is a random sample of the regression process.

Stone (1977) studied estimation problem of conditional quantiles using nonparametric regression and suggested estimating conditional quantiles by quantiles of nonparametric weighted conditional distribution function:

$$\hat{F}_n(y|X) = \sum_i W_{ni}(X) I(Y_i \leq y) \tag{1.7}$$

where $W_{ni}(X)$ are weights reflecting the use of information on X.

Magee, Burbidge and Robb (1991) used these quantile functions to detect the distribution of wealth against age in Canada.

For fixed design, Cheng (1983) showed that the asymptotic distribution of the estimator (1.7) is normal taking Stone's weight function W(x) as a probability density function. For random design, Stute (1986) proved a kind of NN-type estimator of conditional quantiles is asymptotic normal by writing Stone's weight function as

$$\frac{K\left(\frac{F_n(X)-F_n(X_i)}{a_n}\right)}{\sum_{i=1}^{n} K\left(\frac{F_n(X)-F_n(X_i)}{a_n}\right)},$$

where

$$F_n(x) = n^{-1} \sum_{i=1}^{n} I(X_i \leq x).$$

Let $X^* = |X - x_0|$ and $\{X^*_{ni}\}_{i=1}^n$ denote the order statistics and $\{Y_{ni}\}_{i=1}^n$ the induced order statistics of $\{(X^*_{ni}, Y_{ni})\}_{i=1}^n$. For any positive integer $k \leq n$, the k-NN empirical cdf of Y (with respect to x_0) is

$$\hat{G}_{nk}(y) = k^{-1} \sum_{i=1}^{k} I(Y_{ni} \leq y) \tag{1.8}$$

Bhattacharya and Gangopadhyay (1990) defined $k - NN$ estimators of p-quantile by p-quantile of $\hat{G}_{nk}$. The kernel estimator with uniform kernel can be expressed as

$$inf\{z : \hat{G}_{nk}(y) \geq [n\hat{F}_n(h/2)p]/n\hat{F}_n(h/2)\} \tag{1.9}$$

with $\hat{F}_n$ being the empirical cdf of $\{X^*_{ni}\}_{i=1}^n$, where $[u]$ means "integer part" of u. Which leads to Bahadur-type representations of estimators.

Jones and Hall (1990) made a good theoretical MSE investigation based on "check function" and gave a kernel estimate $\hat{q}_p(x)$ of $q_p(x)$ as the solution θ of the following equation based on $\{X_i, Y_i\}_1^n$,

$$H_p(\theta) \equiv \sum_{i=1}^{n} W_i(x)\psi(Y_i - \theta) = 0,$$

where

$$\psi(z) = pI_{(0,\infty)}(z) - (1-p)I_{(-\infty,0)}(z), \quad z \neq 0,$$

is the derivative of $\rho(z)$, except for being undefined at $z = 0$. Their work was extended to local linear fitting by Fan, Hu and Truong (1994).

Chaudhuri (1991) defined an estimate of $q_p(x)$ as the solution β which minimizes

$$\sum \rho_p(Y_i - P_n(\beta, X_i)),$$

where $P_n(\beta, x) = \sum \beta_u n^{-1/2a[u]} x^u$.

Quantile regression, particularly the approach in terms of check function, has been received much attention in the econometric literature.

1.2.2 Smoothing Reference Charts in Medicine

There are two main features of medical reference charts. Firstly, the covariate concerned (often age) is continuous, and the measurements, such as height, weight or middle-upper-arm-circumference have continuous distributions as well. This fact is taken into consideration when constituting reference charts. Secondly, methods for fitting smooth reference charts have been proposed in the last few

years. Most of biometrical measurements are assumed to have approximately normal distribution. However, some measurements, such as weight, circumferences and skinfolds have distributions at a fixed age which are often non-Gaussian, and it is more appropriate not to make strong distributional assumptions.

Suggestions of whether to fit parametric, nonparametric or semi-parametric models basically follow two lines:

(i) Look for a suitable transformation, such as Box-Cox power transformation, logarithms or Johnson transformation.

(ii) Assume that centiles curves can be fitted by parametric forms, such as a polynomial.

However, only nonparametric way has maximum flexibility on the model and distribution assumptions. Applications of smoothing conditional quantiles -smoothing reference charts in medicine using nonparametric techniques are discussed by Cole (1988) and Healy, Rasbash and Yang (1988), also Rossiter (1991), Cole and Green (1992).

Cole's LMS method uses different transformations for each level of the time (age) variable, chosen from Box and Cox family. Smooth curves are fitted to the estimated centiles by smoothing the maximum likelihood estimates of the median (M), standard deviation (S) and power index across time (L). There are various ways in which the above three parameters could be estimated from data. Green (1988) proposed a single-stage fitting procedure for three parameters based on penalized likelihood estimation.

Jones (1988) in a discussion of Cole's paper suggested "spline smoothing regres-

sion quantile" by considering minimizing

$$R_p(f) = \sum_1^n \rho_p(Y_j - f(t_j)) + \lambda \int [f''(t)]^2 dt.$$

Essentially, this is a "check function", but with roughness penalty instead of kernel smoothing .

A kernel-type estimator can be obtained (Rossiter, 1991) by first estimating the bivariate density by a bivariate kernel such as bivariate logistic density function, then integrating the resulting joint density function to obtain the conditional density functions and hence the conditional distribution function.

Healy, Rasbash and Yang (1988) used a method based on the technique for smoothing a scatter diagram described by Cleveland (1979) that is by fitting a polynomial quantile function with polynomial coefficients related to normal centile points (see Chapter 6). The order of fitting polynomial is an important factor here.

Wright (1995) recently discussed the application of quantiles smoothing for censored data in survival analysis in her Ph.D thesis.

1.3 Motivation to Use Nonparametric Smooth for Regression Quantiles

Generally, regression is dictated by the model when applying parametric regression, and by the data for nonparametric regression. Even when parametric modelling is the ultimate goal, nonparametric methods can prove useful for

exploratory analysis and for checking and improving functional models. This flexibility has more particular meaning in regression quantiles than that of general regression contexts.

1.3.1 Flexibility for Non-Gaussianity of Data

Gaussian assumption for model or data is hardly reasonable in practice although it is sometimes argued on general grounds that certain type of measurements, e.g. biometric data, are approximately normally distributed.

For example, height is an essential measurement for monitoring somatic growth and development. The method usually recommended for the construction of age-related centile charts for height involves calculating the mean and standard deviation of the measurement at each of a sequence of ages, and these are smoothed either graphically or by curve-fitting, then the centiles are calculated by

$q_p(x)$ =mean(x)+standard deviation (x)×the p-centile of standard normal.

However, this is less satisfactory for weight, circumferences and skinfolds, which have often non-Gaussian distribution. Usually some transformation (Box-Cox transformation, log-transformation and Johnson-transformation) are introduced to attain Gaussianity (Van't Hof, Wit & Roede, 1985, Cole, 1988, Thompson & Theron, 1990, Royston, 1991). Obviously, the maximum flexibility for this will be free-distribution assumption.

1.3.2 Flexibility for Modelling

The link using quantile function between response and covariate is complicated and changes from data to data. Without any prior information, we have seen that even a good polynomial-parametric form does not always give satisfactory results (Pan, Goldstein & Yang, 1991).

1.4 The Data

Developing practical methods with good theoretical support is the fundamental goal of this research. To address the practical performance of each method, data sets are simulated and analyzed accordingly. Also, these methods are applied using five different sets of data from the literature arising from some practical experiments. The scatter plots of these data are shown in Figures 1.1-1.5.

The first data set comes from an anthropometry survey on triceps skinfold of 892 girls and women up to age 50 in three Gambian villages, seen during the dry season of 1989, and shown in figure 1.1. Note the "notch" in the dependence at around 9 years, and apparently positively skewed distribution for ages less than 20. Also, 620 (70 per cent) of the subjects were aged under 20. Cole & Green (1992) used a different method to calculate seven quantiles for these data.

The second data set (figure 1.2) comprises of serum concentration (grams per litre) of immunoglobulin–G in children age 6 months to 6 years. Royston & Altman (1994) used it with $n = 298$ but the data kindly sent by Dr Royston contained $n = 300$ points, all of which we use. The data came originally from Isaacs et al (1983) who "aimed to establish reference centiles for the serum concentra-

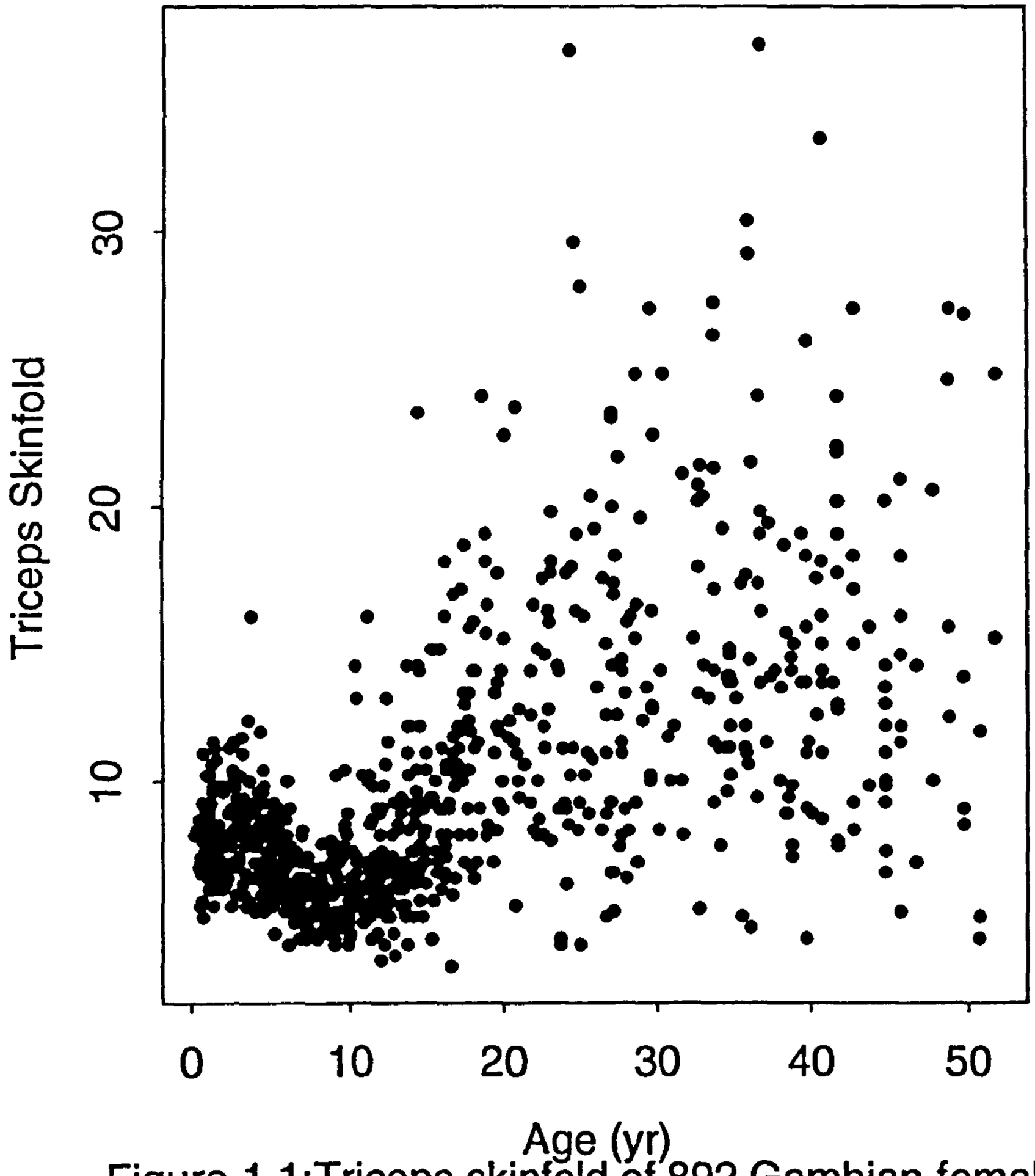


Figure 1.1:Triceps skinfold of 892 Gambian females

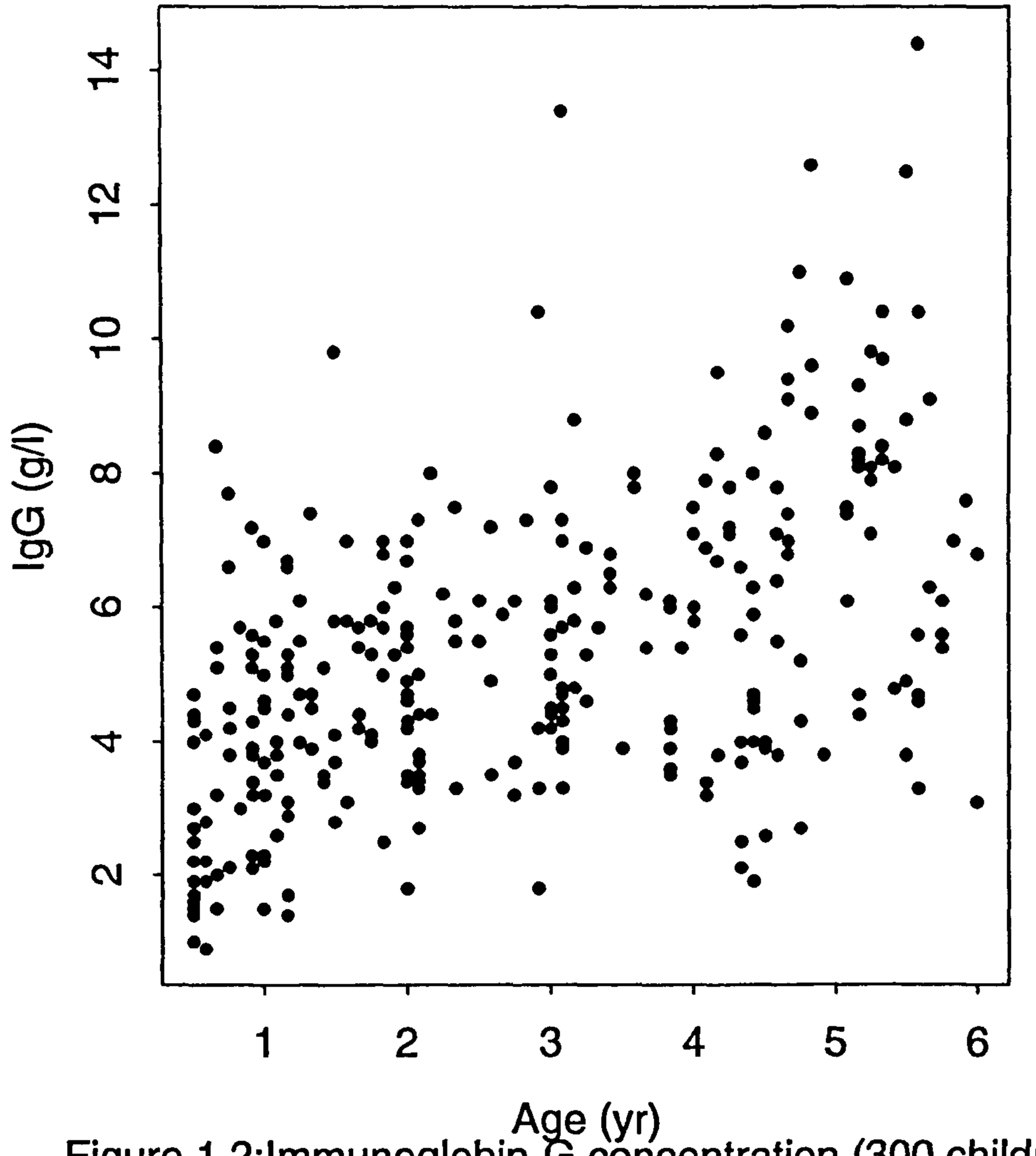


Figure 1.2:Immunoglobin-G concentration (300 children)

tion of certain immunoglobins in children". The square-root transformation of this data is almost normal.

The third set of data (see figure 1.3) consists of 184 survival time of heart transplant patients of age 12-64 years based on the Stanford heart transplant survey originally described by Crowley & Hu (1977). Associate with each patient is a date of acceptance and a date last seen, and the survival time is defined as the difference of two dates. This version of the data is that used (possibly without taking logs of responses) by Ms Eileen Wright as part of her Ph.D. studies.

The fourth example was also used by Cole and Green (1992), obtained as part of the American HANESI Health and Nutrition Survey. It consists of body weight of 4011 U.S. girls aged between 1 and 21 years. The distribution is skew, with a steady increase in the earliest period, then acceleration until 12 years then a slowing down.

The final set of data is the famous motorcycle impact data (e.g. Härdle, 1990) which comprises accelerometer readings against times in an experiment on the efficacy of motorcycle helmets. Like many other authors, we disregard here the potential for treating the points as a time series. The data set contains 133 points. Koenker, Portnoy & Ng (1992) analyzed the data by drawing three quantile curves, the 50th, 10th and 90th, using a different smoothing method.

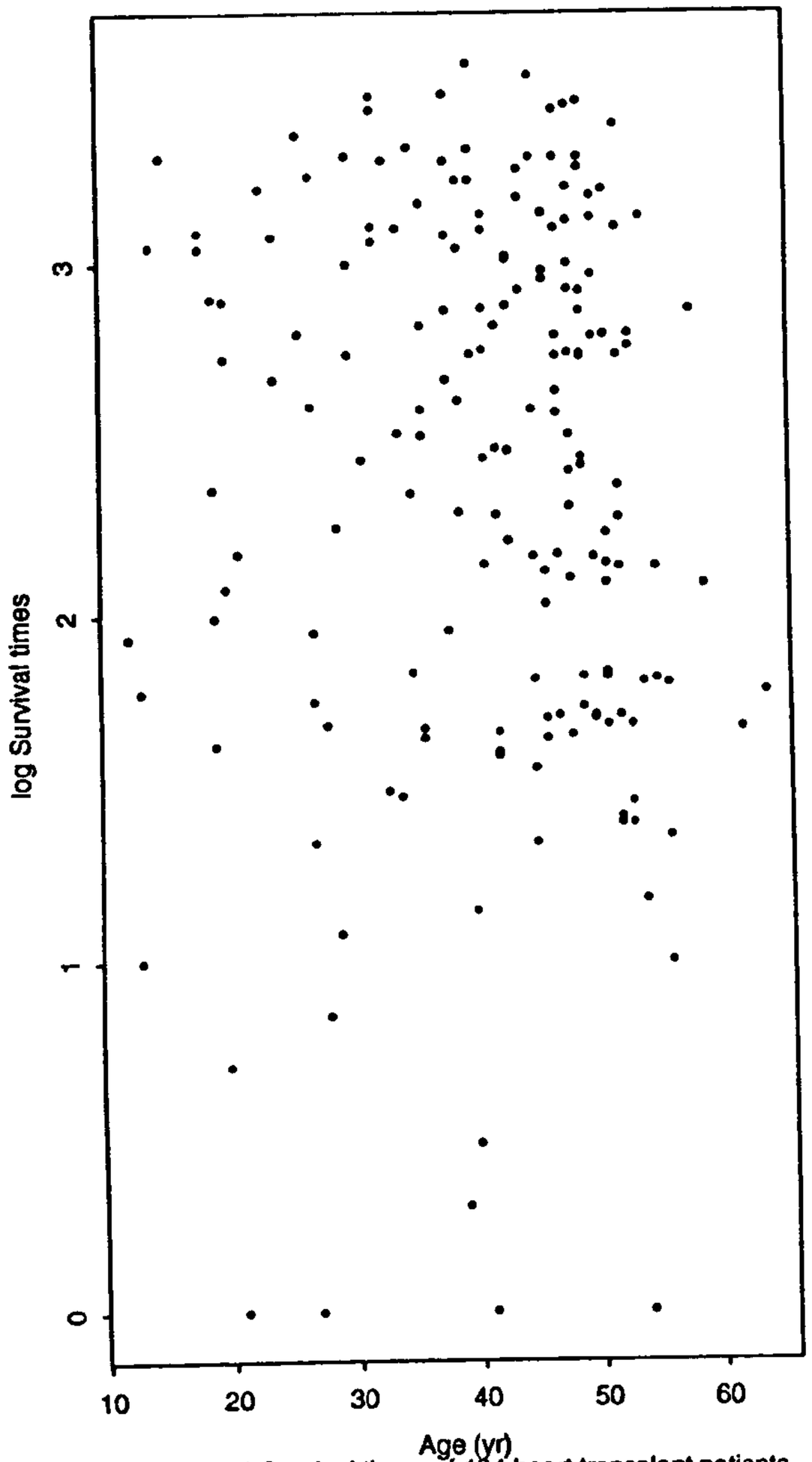


Figure 1.3:Survival times of 184 heart transplant patients

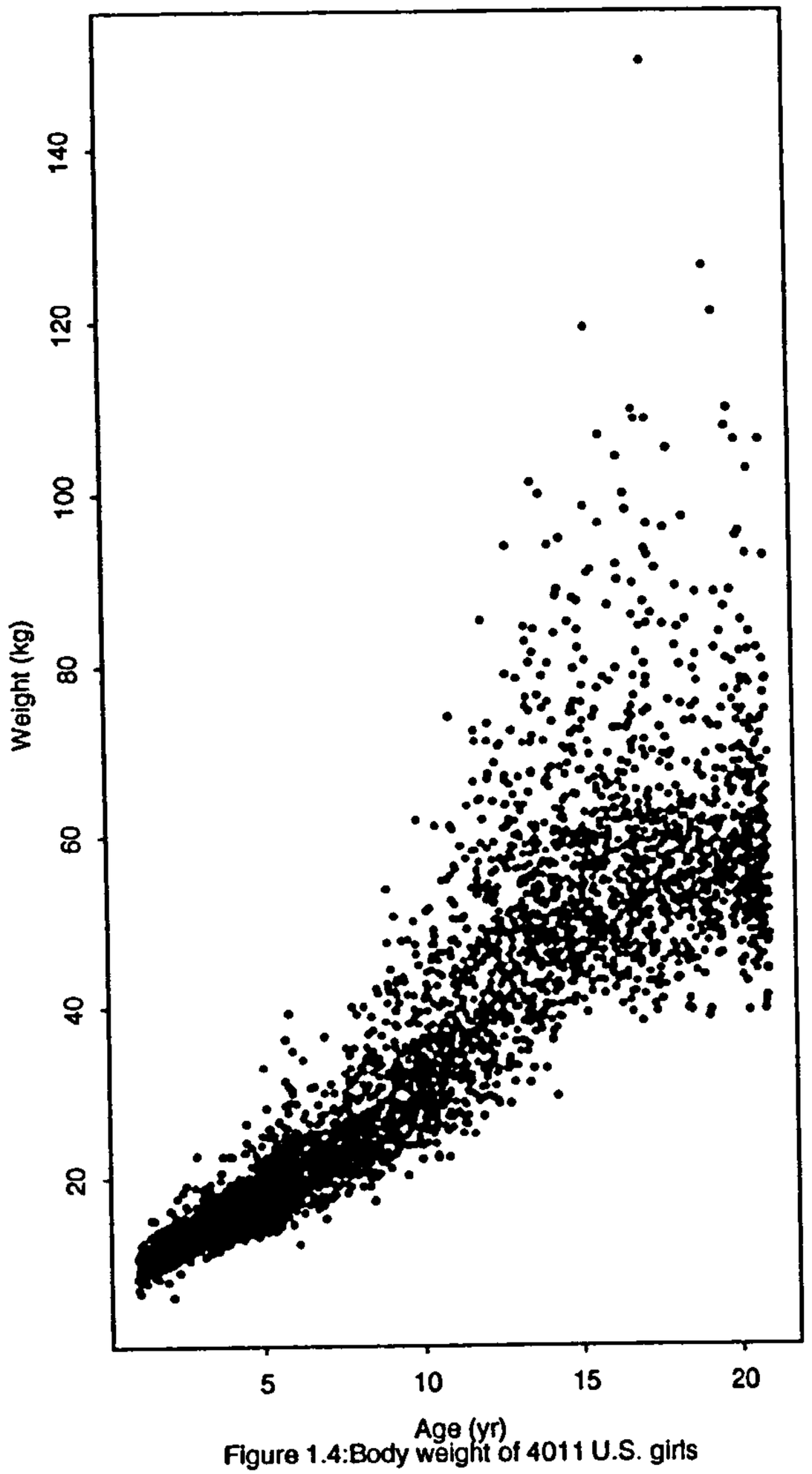


Figure 1.4:Body weight of 4011 U.S. girls

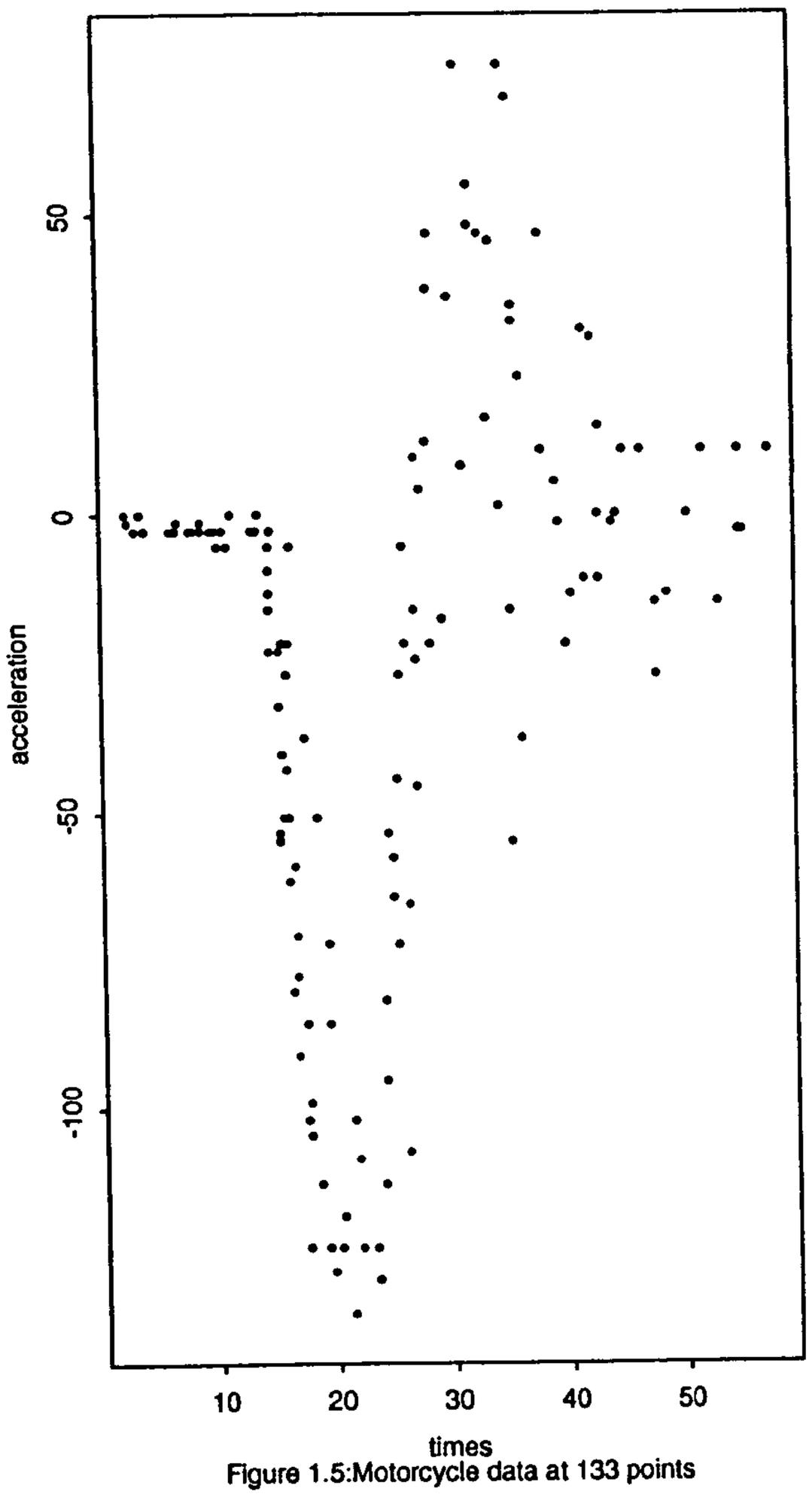


Figure 1.5:Motorcycle data at 133 points

Chapter 2

Local Linear Kernel Fitting Regression Quantiles

2.1 Introduction

As mentioned in Chapter 1, the seminal (parametric) work of Koenker & Bassett (1978) was a major step in estimating regression quantiles. In the current chapter, we are concerned with the nonparametric estimation of regression quantile functions. For estimation of the regression mean, local polynomial fitting, particularly its special case local linear fitting, have become increasingly popular. This is the stuff of loess (Stone, 1977, Cleveland, 1979, Cleveland & Devlin, 1988) and it has also been further recognised to have various advantages through recent work such as Fan (1992), Fan & Gijbels (1992, 1995), Hastie & Loader (1993), Ruppert & Wand (1994) and Cleveland & Loader (1995); see also Wand & Jones (1995, Chapter 5). Unsurprisingly, local polynomial fitting, and in particular local linear fitting, can be adapted to quantile regression and its advantages will

be maintained, as will be described later.

The purpose of this chapter is to develop local linear approaches to quantile regression in such a way that the results are immediately applicable in practice. The basic techniques are not novel (Chaudhuri, 1991, Fan, Hu & Truong, 1994, and Fan, Yao & Tong, 1996, provide relevant theoretical background) but many details and insights are.

In fact, we present and develop two alternative local linear quantile regression methods which, as we shall see on a variety of real data examples in Section 2.4, are broadly equivalent in terms of results. While the user could certainly contemplate employing either, we will show that we have a preference for the second one we describe. The first method is, however, more direct. The estimated quantile function $\hat{q}_p(x)$ is based on minimizing a local linear kernel weighted version of $E\{\rho_p(Y-a)|X=x\}$ where ρ_p is the "check" function given by (1.3)

$$\rho_p(z) = pzI_{[0,\infty)} - (1-p)zI_{(-\infty,0)}(z) \tag{2.1}$$

where as before p indexes the conditional quantile of current interest. This method involves a kernel localisation function K which we take to be a symmetric probability density function, and its scale parameter h is the bandwidth that controls the amount of smoothing applied to the data. Motivation, implementation, and asymptotic mean squared error properties are discussed in Section 2.2.1 to 2.2.3. Section 2.2.4 deals with a novel bandwidth selection rule covering all desired conditional quantiles which is fully practical and is successfully used in our examples.

In Section 2.3, a parallel development is made of an alternative "double kernel" approach to conditional quantile estimation by taking a kernel weighted local linear approach to estimating the conditional distribution function. In this case,

two bandwidths are allowed, an additional one "in the y direction" added to the existing one "in the x direction". Having to select a second bandwidth is an unappealing feature, but in Section 2.3.4, we show how to specify a rule–of–thumb for it that works in conjunction with the rule of Section 2.2.4 for the x direction bandwidth. The theoretical work of Section 2.3.3, plus practical experience, shows that the precise value of this second bandwidth is, unsurprisingly, not critical, but a practical rule (which depends on p) is still needed and is given here. To define the double kernel quantile estimator $\tilde{q}_p$, solve

$$p = \frac{1}{\sum_j w_j(x)} \sum_j w_j(x) \Omega(\frac{\tilde{q}_p(x) - Y_j}{h_2}). \tag{2.2}$$

where Ω is the distribution function associated with a kernel function W. Here, $w_j(x)$ is the weight function associated with local linear fitting:

$$w_j(x) = K(\frac{x - X_j}{h_1})[S_{n,2} - (x - X_j)S_{n,1}] \tag{2.3}$$

with

$$S_{n,l} = \sum_{i=1}^{n} K(\frac{x - X_j}{h_1})(x - X_i)^l, \quad l = 1, 2,$$

and the two bandwidths h_1, h_2 relate respectively to x and y direction smoothing. The right–hand side of (2.2) is a (double kernel) local linear estimate of the conditional distribution function, and the equation defines its inverse, our conditional quantile estimator. Further details on $\tilde{q}_p$ are in Section 2.3.1 and implementation details in Section 2.3.2.

The preference for $\tilde{q}_p$ over $\hat{q}_p$ resides in the smoother appearance of the former and a much reduced propensity for estimated quantiles to cross. A simulation about comparing $\tilde{q}_p$ and $\hat{q}_p$ and other estimators will be carried out in Chapter 7. In general, $\tilde{q}_p$ performs better than $\hat{q}_p$ in that study.

This chapter forms the basis of the paper Yu and Jones (1997a).

2.2 Local Linear Check Function

2.2.1 The Method

Suppose that $(X_1, Y_1), ..., (X_n, Y_n)$ is a set of independent observations from some underlying distribution $F(x, y)$ with density $f(x, y)$. Interest centres on the responses Y_i which are considered to be realisations from a conditional distribution $F(y|x)$ or density $f(y|x)$ of Y given $X = x$. A characterisation of the pth conditional quantile $q_p(x)$ is

$$q_p(x) = \text{argmin}_a E\{\rho_p(Y - a)|X = x\}$$

with ρ_p is given by (2.1). A first, "local constant", sample version of this might be defined by

$$\bar{q}_p(x) = \text{argmin}_a \sum_{i=1}^{n} \rho_p(Y_i - a)K(\frac{x - X_i}{h}).$$

Here h and K are respectively the bandwidth and kernel function.

In regression mean estimation, however, local linear fitting is nowadays considered superior to local constant fitting (e.g Wand & Jones, 1995). A direct comparison between local constant and local linear approaches in the conditional quantile estimation setting is made in Chapter 3. The idea of the local linear fit is to approximate the unknown pth quantile q_p by a linear function

$$q_p(z) = q_p(x) + q_p'(x)(z - x) \equiv a + b(z - x)$$

for z in a neighborhood of x. Locally, estimating $q_p(x)$ is equivalent to estimating a while estimating $q_p'(x)$ is equivalent to estimating b. This motivates us to define an estimator by setting $\hat{q}_p(x) = \hat{a}$, where $\hat{a}$ and $\hat{b}$ minimize

$$\sum_{i=1}^{n} \rho_p(Y_i - a - b(X_i - x))K(\frac{x - X_i}{h}). \tag{2.4}$$

(See Chaudhuri (1991), Koenker, Portnoy & Ng (1992) and Fan, Hu & Truong (1994)). This method maintains various advantages of local linear mean fitting, such as design adaptation and good boundary behaviour, in a conditional quantile context.

2.2.2 The algorithm

A difficulty in calculating the kernel smoothing resulting from the check function $\rho_p(z)$ is that $\hat{q}_p(x)$ does not have an explicit representation. In addition, the derivatives of $\rho_p(z)$ do not exist everywhere, and there are two parameters to minimize over. To solve the minimizing problem (2.4), either approximate $\rho_p(z)$ by a smooth function or optimize the function $\rho_p(z)$ using an algorithm and software development, as no ready–made software is available for solving this problem. We should stress that this is just an algorithm that works rather than one which is necessarily "optimal".

The approach is that of iteratively reweighted least squares (also suggested in Lejeune & Sarda, 1988). Define new weights

$$w_p(x; X_i, Y_i; a, b) = \begin{cases} \frac{p}{Y_i - a - b(X_i - x)} & \text{if } Y_i - a - b(X_i - x) > 0 \\ \frac{p-1}{Y_i - a - b(X_i - x)} & \text{if } Y_i - a - b(X_i - x) < 0 \\ 0 & \text{if } Y_i - a - b(X_i - x) = 0 \end{cases}$$

$$K_p(x; X_i, Y_i; a, b) = w_p(x; X_i, Y_i; a, b) K(\frac{x - X_i}{h})$$

then

$$(\hat{a}, \hat{b}) = \text{argmin}_{(a,b)} \sum_{i=1}^{n} (Y_i - a - b(X_i - x))^2 K_p(x; X_i, Y_i; a, b).$$

Initially use a guess (a_0, b_0) then use the above formulation to iterate to convergence. In particular, if (a_l, b_l) are the values of (a, b) at the lth iteration, the next values (a_{l+1}, b_{l+1}) will be given by

$$a_{l+1} = \frac{\sum_j K_p(x; X_j, Y_j; a_l, b_l)\Big(T_{n,2}(a_l, b_l) - (x - X_j)T_{n,1}(a_l, b_l)\Big)Y_j}{T_{n,0}(a_l, b_l)T_{n,2}(a_l, b_l) - T_{n,1}^2(a_l, b_l)}$$

$$b_{l+1} = \frac{\sum_j K_p(x; X_j, Y_j; a_l, b_l)\Big((x - X_j)T_{n,0}(a_l, b_l) - T_{n,1}(a_l, b_l)\Big)Y_j}{T_{n,0}(a_l, b_l)T_{n,2}(a_l, b_l) - T_{n,1}^2(a_l, b_l)}$$

Here,

$$T_{n,l}(a, b) = \sum_j K_p(x; X_j, Y_j; a, b)(x - X_j)^l, \qquad l = 0, 1, 2.$$

Note that the above is a pointwise algorithm and that an entire curve or at least values on a fine grid is needed. Simply fix each x in turn and then obtain the collection of values of $\hat{q}_p(x)$ for all p of interest. For the first gridpoint x_0 say, the starting points used are as follows: for the median ($p = 0.5$), start with the regression mean estimator $\hat{m}(x_0)$ and its derivative $\hat{m}_1(x_0)$ by the local linear least squares approach i.e.

$$\hat{m}(x_0) = \frac{\sum_j K(\frac{x_0 - X_j}{h})\Big(S_{n,2} - (x_0 - X_j)S_{n,1}\Big)Y_j}{S_{n,0}S_{n,2} - S_{n,1}^2}$$

$$\hat{m}_1(x_0) = \frac{\sum_j K(\frac{x_0 - X_j}{h})\Big((x_0 - X_j)S_{n,0} - S_{n,1}\Big)Y_j}{S_{n,0}S_{n,2} - S_{n,1}^2}.$$

Then, for each $p > 1/2$ in increasing order, use the previously found $\hat{q}(x_0)$ (and its associated $\hat{b}(x_0)$) as initial guess, for example, use $\hat{q}_{0.75}(x_0)$ perhaps as initialiser for finding $\hat{q}_{0.9}(x_0)$. Likewise, work downwards from $(\hat{m}(x_0), \hat{m}_1(x_0))$ for each $p < 1/2$. Then, for the next gridpoint x_1, say, calculate $\hat{q}_p(x_1)$ from starting point $\hat{q}_p(x_0)$.

We found that the convergence rate of this iterative procedure was very fast at the interior points, but a little slower near the boundary. Using the convergence

criterion with stopping rule $|a_{l+1}(x) - a_l(x)| < 10^{-2}\bar{r}$ for each x, where $\bar{r}$ is the mean absolute residual at the first iteration, the typical number of iterations needed was 12.

Note that this algorithm automatically produces derivative curves for the quantiles at the same time as estimating the quantiles themselves, albeit using the same value of the smoothing parameter for both. Also, a convergence theorem can be found in Lejeune & Sarda (1988).

2.2.3 Mean Squared Error of $\hat{q}_p(x)$

To assess the performance of $\hat{q}_p(x)$ we evaluate its (conditional or unconditional) mean squared error (MSE). For local linear conditional quantile fitting, the asymptotic form of $\mathrm{MSE}(\hat{q}_p(x))$ (as $n \to \infty$, $h = h(n) \to 0$, $nh \to \infty$ and under certain conditions) is given by Fan, Hu & Truong (1994) as:

$$\mathrm{MSE}(\hat{q}_p(x)) \simeq \frac{1}{4}h^4\mu_2(K)^2 q_p''(x)^2 + \frac{R(K)p(1-p)}{nhg(x)f(q_p(x)|x)^2} \tag{2.5}$$

where

$$\begin{aligned} \mu_2(K) &= \int u^2 K(u)du, \\ R(K) &= \int K^2(u)du, \end{aligned}$$

and g is the marginal density of X (design density). Also from Fan, Hu & Truong (1994), if $x = ch$, $0 < c < 1$, is a boundary point of the design, for K to have support $[-1, 1]$ and g to have support $[0, 1]$, then

$$\mathrm{MSE}(\hat{q}_p(ch)) \simeq \frac{1}{4}h^4\alpha_c^2(K)q_p''(0+)^2 + \frac{\beta_c(K)p(1-p)}{nhg(0+)f(q_p(0+)|0+)^2}$$

where

$$\alpha_c(K) = \frac{a_2^2(c;K) - a_1(c;K)a_3(c;K)}{a_0(c;K)a_2(c;K) - a_1^2(c;K)}, \qquad \beta_c(K) = \frac{\int_{-1}^{c}\{a_2^2(c;K) - a_1(c;K)u\}^2 K(u)du}{a_0(c;K)a_2(c;K) - a_1^2(c;K)},$$

$$a_l(c;K) = \int_{-1}^{c} u^l K(u)du, \quad l = 0, 1, 2.$$

and $g(0+) = \lim_{z \downarrow 0} g(z)$.

Major advantages of local linear fitting, as reflected by MSE are (i) no dependence of the asymptotic bias on the design density g, and indeed its dependence only on the simple quantile curvature function q''_p, and (ii) automatic good behaviour at boundaries, at least with regard to orders of magnitude, and no need for further boundary correction.

2.2.4 Bandwidth selection

With the basic model in place, one has to face the important bandwidth selection problem. Since the quality of the curve estimates depends sensitively on the choice of h, a convenient and effective data–based rule will always be needed. However, almost nothing has been done so far about this problem in the context of estimating $q_p(x)$, and the selection problem is difficult even in the simpler case of kernel density estimation (Jones, Marron & Sheather, 1996).

The asymptotically optimal (interior) bandwidth h_p is used as a starting point where

$$h_p^5 = \frac{R(K)p(1-p)}{n\mu_2(K)^2 q''_p(x)^2 g(x) f(q_p(x)|x)^2}. \tag{2.6}$$

This gives a relationship between optimal bandwidths for different values of p:

$$\left(\frac{h_{p_1}}{h_{p_2}}\right)^5 = \frac{p_1(1-p_1)}{p_2(1-p_2)} \frac{q''_{p_2}(x)^2 f(q_{p_2}(x)|x)^2}{q''_{p_1}(x)^2 f(q_{p_1}(x)|x)^2} \tag{2.7}$$

Though $q_p(x)$ itself might vary considerably with x in terms of curvature at any one point the second derivatives of any two quantiles will often be very

similar. For example, following the usual type of parametric regression model with identically distributed errors, the quantiles would be parallel and hence their second derivatives are equal. More generally, it seems reasonable as a first order approximation to take $q''_{p_1}(x) = q''_{p_2}(x)$. Equality is not an appropriate approximation for $f(q_p(x)|x)$ since this should be different for rather different p, however we can turn to the usual type of "rule–of–thumb" calculations based on assuming a normal (conditional) distribution at this stage as an appropriate approach. Suppose for a moment that f is the density of a normal distribution with mean μ_x and variance σ_x^2. Then if $Z_p = \Phi^{-1}(p)$ denotes the pth quantile of the standard normal distribution, $f(q_p(x)|x) = \sigma_x^{-1}\phi(Z_p)$, and

$$f(q_{p_2}(x)|x)/f(q_{p_1}(x)|x) = \phi(\Phi^{-1}(p_2))/\phi(\Phi^{-1}(p_1))$$

where ϕ and Φ are standard normal density and distribution functions respectively.

Employing these approximations in (2.7) yields

$$\left(\frac{h_{p_1}}{h_{p_2}}\right)^5 = \frac{p_1(1-p_1)}{p_2(1-p_2)}\frac{\phi(\Phi^{-1}(p_2))^2}{\phi(\Phi^{-1}(p_1))^2}$$

which gives a neat, explicit and practical way of modifying h with p.

In particular, when $p_2 = 1/2$,

$$h_p^5 = \pi^{-1}2p(1-p)\phi(\Phi^{-1}(p))^{-2}h_{1/2}^5.$$

It remains to find a method of selection of a bandwidth for the median. In fact, the automatic bandwidth $h_{1/2}$ can be expressed in terms of h_{mean} (the optimal choice of h for regression mean estimation) which has already been considered elsewhere (Fan & Gijbels, 1995, Ruppert, Sheather & Wand, 1995). From Fan (1993),

$$h_{mean}^5 = \frac{R(K)\sigma^2(x)}{n\mu_2(K)^2\{m''(x)\}^2 g(x)}$$

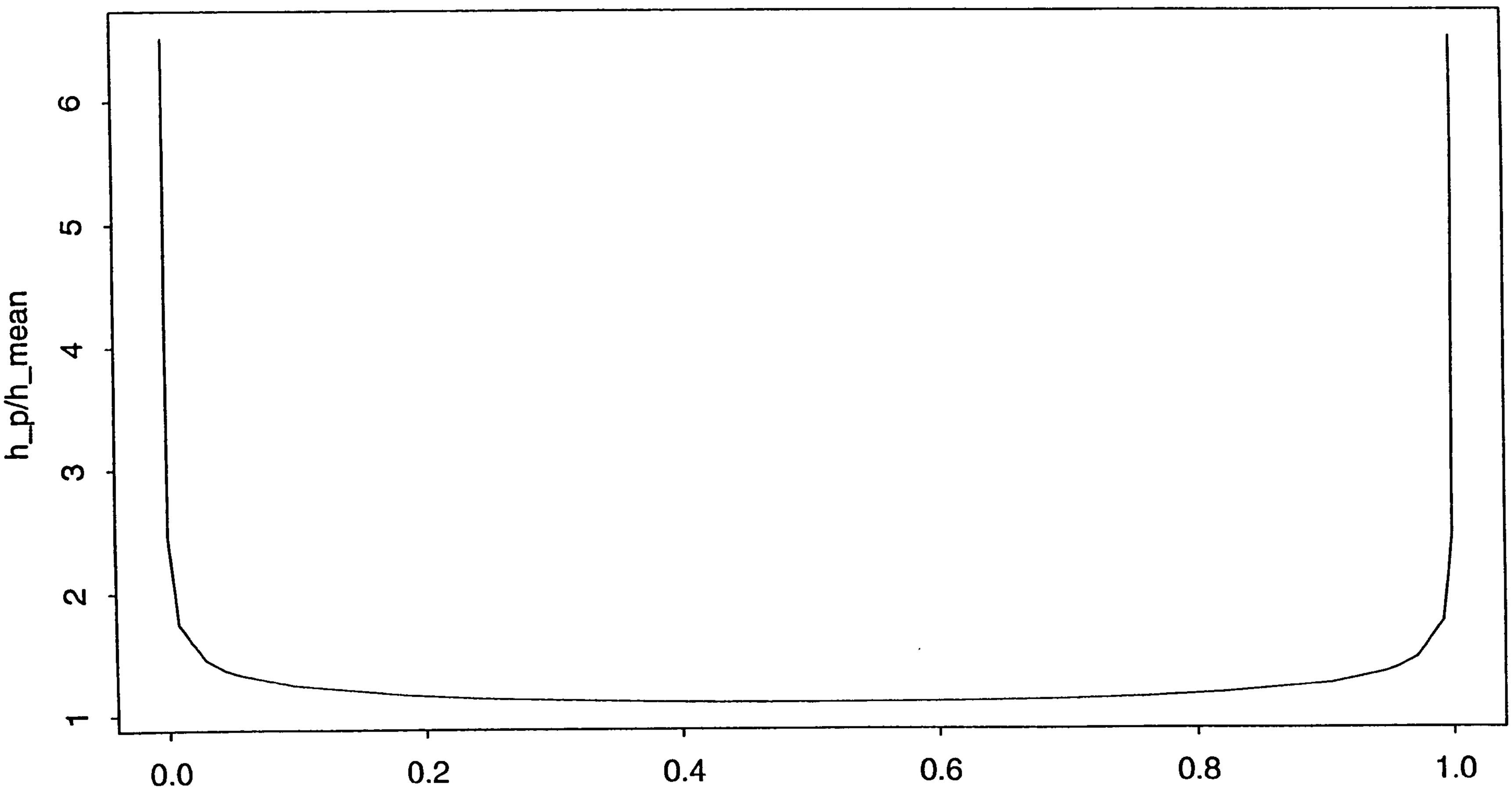


Figure 2.1: Relationship between h_p/h_mean against p by rule-of-thumb

where $m(x)$ and $\sigma^2(x)$ are the conditional mean and variance respectively. It follows that

$$\left(\frac{h_{mean}}{h_{1/2}}\right)^5 = \frac{4q''_{1/2}(x)^2\sigma^2(x)f(q_{1/2}(x)|x)^2}{m''(x)^2}.$$

By the same arguments as above, $q''_p(x)$ and $m''(x)$ are similar and may be set equal, and employ the normal distribution to argue that $\sigma^2(x)f(q_{1/2}(x)|x)^2$ be replaced by $\phi(\Phi^{-1}(1/2))^2 = (2\pi)^{-1}$. Thus

$$\left(\frac{h_{mean}}{h_{1/2}}\right)^5 = \frac{2}{\pi}$$

Summarizing, the automatic bandwidth selection strategy for smoothing conditional quantiles as follows:

(a) use ready–made and sophisticated methods to select h_{mean}, e.g. employ Ruppert, Sheather & Wand (1995) technique,

(b) use $h_p = h_{mean}\left\{\frac{p(1-p)}{\phi(\Phi^{-1}(p))^2}\right\}^{1/5}$ to obtain all other h_ps from h_{mean}.

For different quantiles, the quantity $\left\{\frac{p(1-p)}{\phi(\Phi^{-1}(p))^2}\right\}^{1/5}$ is plotted against p in Figure 2.1, and the selected values using rule-of-thumb are displayed in Table 2.1. These show the expected minimal smoothing for the median and a gradually increasing smoothing for the less central quantiles. The increase is, of course, symmetric in p and $1-p$. However, the change is very small for moderate p and only becomes considerable as p gets close to 0 or 1 (in fact, $\lim_{p\to 0} b(p) = \lim_{p\to 0} b(p) = \infty$ if let $b(p) = \left\{\frac{p(1-p)}{\phi(\Phi^{-1}(p))^2}\right\}^{1/5}$).

p	h
0.025 or 0.975	$1.48h_{mean}$
0.03 or 0.97	$1.44h_{mean}$
0.05 or 0.95	$1.34h_{mean}$
0.1 or 0.9	$1.24h_{mean}$
0.25 or 0.75	$1.13h_{mean}$
0.5	$1.095h_{mean}$

Table 2.1: The relationship between h_p and h_{mean}

2.3 Double kernel smoothing

2.3.1 The Method

Given a symmetric kernel W, let

$$\Omega(y) = \int_{-\infty}^{y} W(u)du.$$

Note that

$$\int_{-\infty}^{y} W_{h_2}(Y_j - u)du = \Omega(\frac{y - Y_j}{h_2}).$$

Usually, W is a density and Ω is its distribution function.

Then, as $h_2 \to 0$,

$$E\{\Omega(\frac{y - Y}{h_2})|X = x\} \approx F(y|x).$$

Moreover for local linear approach, further approximation leads to

$$\begin{aligned} E\{\Omega(\frac{y - Y}{h_2})|X = z\} &\approx F(y|z) \\ &\approx F(y|x) + \dot{F}(y|x)(z - x) \\ &\equiv a + b(z - x) \end{aligned}$$

where $\dot{F}(y|x) = \partial F(y|x)/\partial x$.

Define $\tilde{F}_{h_1,h_2}(y|x) = \tilde{a}$ where

$$(\tilde{a}, \tilde{b}) = \text{argmin} \sum_i \left(\Omega(\frac{y - Y_i}{h_2}) - a - b(X_i - x) \right)^2 K(\frac{x - X_i}{h_1}).$$

This *conditional distribution function estimate* is closely related to the conditional density function estimate of Fan, Yao & Tong (1996).

Finally, for conditional quantile estimation, define $\tilde{q}_p(x)$ as the solution of

$$\tilde{F}_{h_1,h_2}(\tilde{q}_p(x)|x) = p$$

or

$$\tilde{q}_p(x) = \tilde{F}^{-1}_{h_1,h_2}(p|x) \tag{2.8}$$

with

$$\tilde{F}_{h_1,h_2}(\tilde{q}_p(x)|x) = \frac{1}{\sum_j w_j(x; h_1)} \sum_j w_j(x; h_1) \Omega(\frac{\tilde{q}_p(x) - Y_j}{h_2}) \tag{2.9}$$

where the weights $w_j(x; h_1)$ are precisely $w_j(x)$ in (2.3) with h renamed h_1.

This alternative approach, via the conditional distribution function, is also attractive, but suffers from the disadvantage of having to specify a second bandwidth h_2 as well as the bandwidth h_1 which plays much the same role as the bandwidth h in Section 2.2.4. Unsurprisingly, it turns out that the estimates are not particularly sensitive to the value of h_2. The choice $h_2 = 0$ is, however, not attractive since it results in a discontinuous conditional quantile estimate (which could be smoothed once more, but the approach is inelegant).

2.3.2 The Algorithm

Two kernel functions K and W have to be specified. At this point we need not be specific about K since it has the same role as K in Section 2.2.1. Relatively, h_2 has small effect and consequently there is a relatively small effect of h_2, there is a small effect of W, therefore choose W as the uniform kernel, $W(u) = 1/2I(|u| \leq 1)$, for the consideration of simple calculation. This affords

$$\begin{aligned}
\int_{-\infty}^{Q} W_{h_2}(Y_j - u)du &= 1/(2h_2)\int_{-\infty}^{Q} I(Y_j - h_2 \leq u \leq Y_j + h_2)du \\
&= 1/(2h_2)I(Q \geq Y_j - h_2)\int_{Y_j - h_2}^{min(Q, Y_j + h_2)} du \\
&= 1/(2h_2)\left\{(Q - Y_j + h_2)\right. \\
&- (Q - Y_j - h_2)I(Q \geq Y_j + h_2) \\
&- \left.(Q - Y_j + h_2)I(Q \leq Y_j - h_2)\right\}.
\end{aligned}$$

and that $\tilde{q}_p(x)$, for each x, is the value of $Q(x)$ which is the solution of :

$$\begin{aligned}
Q(x) &= \hat{m}(x) + (2p-1)h_2 \\
&+ \frac{\sum_j (Q(x) - Y_j - h_2)w_j(x)I(Y_j \leq Q(x) - h_2)}{\sum_j w_j(x)} \\
&+ \frac{\sum_j (Q(x) - Y_j + h_2)w_j(x)I(Y_j \geq Q(x) + h_2)}{\sum_j w_j(x)}
\end{aligned} \tag{2.10}$$

where $\hat{m}(t)$ is the local linear kernel estimator $\{\sum_j w_j(x)\}^{-1}\{\sum_j w_j(x)Y_j\}$, of the regression mean. An algorithm is developed by solving the equation (2.10) iteratively for $Q(x)$.

To solve (2.10) iteratively, put an initial value into the right–hand side and evaluate solving the equation to give a new value, then repeat the process. The starting point and convergence criterion here are the same as those for computing $\hat{q}_p(x)$. The typical number of iterations needed is 8, which is less than what is required in the other method.

There are many signs that the iterative method works well though have been unable to prove its guaranteed convergence. First, if the weights $w_j(x)$ were non–negative, it is not difficult to show that (2.10) would be a contraction mapping (such non–negativity holds for a local constant version of this method but not necessarily for local linear); without non–negativity, the mapping does not necessarily contract everywhere (but this does not imply non–convergence). Second, as $n \to \infty$, the weights do become non–negative, and the difficulty goes away. And finally, in practice the experience has always been one of convergence to a plausible result.

While it is perfectly possible for $\hat{q}_{p_1}$ and $\hat{q}_{p_2}$ to cross one another (Section 2.4), the same behaviour has not been observed for $\tilde{q}_{p_1}$ and $\tilde{q}_{p_2}$. This is a great attraction of $\tilde{q}$ relative to $\hat{q}$.

2.3.3 The Mean Square Error of $\tilde{q}_p(x)$

To study the properties of estimator $\tilde{q}_p(x)$, particularly MSE, the following conditions are needed, which are based on Fan (1992), and Fan & Gijbels (1995). These conditions are also a suitable specialisation of the requirements of Fan, Hu & Truong, 1994)

(1) The necessary partial derivatives of the joint density and distribution function $f(x, y)$, $F(x, y)$ and of the marginal density $g(x)$ exist and are bounded and continuous at the interior and boundary points.

(2) $g(x) > 0$ and the conditional density $f(y|x) > 0$ are bounded.

(3) The population conditional quantiles $q_p(x)$ are unique.

(4) The two bandwidths have the form: $dn^{-\beta}$, $0 < \beta < 1$.

(5) The kernels W with finite support and K are each second order symmetric.

Also, write

$$F^{ab}(q_p(x)|x) = \frac{\partial^{ab}}{\partial z^a \partial y^b} F(y|z)|_{x,q_p(x)}.$$

Theorem 2.1. For a non-boundary point x, and under Conditions 1-5, if $h_1 \to 0$, $h_2 \to 0$ and $nh_1 \to \infty$, then

$$\begin{aligned} \mathrm{MSE}(\tilde{q}_p(x)) &\simeq 1/4\{\mu_2(K)h_1^2 F^{20}(q_p(x)|x)/f(q_p(x)|x) \\ &+ \mu_2(W)h_2^2 F^{02}(q_p(x)|x)/f(q_p(x)|x)\}^2 \\ &+ \frac{R(K)}{nh_1 g(x) f^2(q_p(x)|x)}\Big(p(1-p) - h_2 f(q_p(x)|x)\alpha(W)\Big) \\ &+ o(h_1^4 + h_2^4 + h_2/nh_1) \end{aligned}$$

where $\alpha(W) = \int \Omega(t)(1-\Omega(t))dt$.

Theorem 2.2 gives $MSE(\tilde{q}_p(x))$ at left boundary points $x = ch_1$, $0 < c < 1$.

Theorem 2.2. Under the conditions of Theorem 2.1, and that $q_p(x)$ is bounded on [0,1] and right continuous at the point 0, then the MSE of the estimator $\tilde{q}_p(x)$ at a boundary point is given by

$$\begin{aligned} \mathrm{MSE}(\tilde{q}_p(ch_1)) &\simeq 1/4\Big\{\alpha_c(K)h_1^2 F^{20}(q_p(0+)|0+)/f(q_p(0+)|0+) \\ &+ \mu_2(W)h_2^2 F^{02}(q_p(0+)|0+)/f(q_p(0+)|0+)\Big\}^2 \\ &+ \frac{\beta_c(K)}{nh_1 g(0+) f^2(q_p(0+)|0+)}\Big(p(1-p) - h_2 f(q_p(0+)|0+)\alpha(W)\Big) \\ &+ o(h_1^4 + h_2^4 + h_2/nh_1). \end{aligned}$$

Proofs of Theorem 2.1 and 2.2:

The following Lemma 1 follows from Theorem 1 of Fan (1993) and Theorem 5 of Fan and Gijbels (1995) (also following the proof of Theorem 5.1 in Fan, Yao and Tong, 1996).

Lemma 1

Let

$$m(x,y) = E\{\Omega(h_2^{-1}(y-Y))|X=x\}$$

and define $\hat{m}_{h_1,h_2}(y|x)$ as the local linear kernel estimator of $m(x,y)$, then under the conditions of Theorem 2.1, as $n \to \infty$,

$$\sqrt{nh_1}\{\hat{m}_{h_1,h_2}(y|x) - m(x,y) - \frac{1}{2}h_1^2\mu_2(K)F^{20}(y|x)\} \to N(0, g^{-1}(x)R(K)\sigma^2(x,y))$$

where $\sigma^2(x,y) = \mathrm{Var}\{\Omega(h_2^{-1}(y-Y))|X=x\}$.

Under the conditions of theorem 2.2, the conditional MSE of the estimator $F_{h_1,h_2}(y|x)$ at the boundary point x_0 is given by

$$1/4\Big(F^{20}(y_0|0+)\alpha_c(K)h_1^2\Big)^2 + \frac{\beta_c(K)}{nh_1 g(0+)}\Big(\sigma^2(0+,y_0)\Big) + o(h_1^4 + 1/nh_1). \quad (2.1)$$

where y_0 denotes the corresponding value of x_0.

Lemma 2

Under the conditions of theorem 2.1,

$$m(x,y) = F(y|x) + 1/2h_2^2\mu_2(W)F^{0,2}(y|x) + O(h_2^2)$$

and

$$\sigma^2(x,y) = F(y|x)(1-F(y|x)) - h_2F^{01}(y|x)\alpha(W) + O(h_2^2)$$

and under the conditions of theorem 2.2,

$$m(0+,y_0) = F(y_0|0+) + 1/2h_2^2\mu_2(W)F^{0,2}(y_0|0+) + O(h_2^2)$$

and

$$\sigma^2(0+, y_0) = F(y_0|0+)(1 - F(y_0|0+)) - h_2 F^{01}(y_0|0+)\alpha(W) + O(h_2^2).$$

Proof. These formulae follow by standard Taylor series approximation, as $h_2 \to 0$, allied to integration–by–parts.

$$\begin{aligned} m(x,y) &= \int_{-\infty}^{+\infty} \Omega(\frac{y-u}{h_2}) f(u|x) \\ &= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{\frac{y-u}{h_2}} W(v)dv \right) f(u|x)du \\ &= 1/h_2 \int_{-\infty}^{+\infty} \left(\int_{u}^{+\infty} W(\frac{y-t}{h_2})dt \right) f(u|x)du \end{aligned}$$

Change the order of integral above, we have

$$\begin{aligned} m(x,y) &= 1/h_2 \int_{-\infty}^{+\infty} W(\frac{y-t}{h_2})dt \int_{-\infty}^{t} f(u|x)du \\ &= \int_{-\infty}^{+\infty} W(t)dt \int_{-\infty}^{y-h_2 t} f(u|x)du \\ &= \int_{-\infty}^{+\infty} W(t)dt \int_{-\infty}^{y} f(v-h_2 t|x)dv \\ &= F(y|x) + 1/2 h_2^2 \mu_2(W) F^{0,2}(y|x) + O(h_2^2) \end{aligned}$$

Further, it has been pointed out that the choice of W is not critical, at least estimating a density function. Consider kernel W with finite range $(-\delta, \delta)$ and $F(y|x)$ given $X = x$ locally regular in $(y - \delta h_2, y + \delta h_2)$, then

$$\begin{aligned} E\{\Omega^2(\frac{y-Y}{h_2})|X = x\} &= \int_{y-\delta h_2}^{y+\delta h_2} \{\Omega(\frac{y-v}{h_2})\}^2 dv + \int_{-\infty}^{y-\delta h_2} f(v|x)dv \\ E\{\Omega(\frac{y-Y}{h_2})|X = x\} &= \int_{y-\delta h_2}^{y+\delta h_2} \Omega(\frac{y-v}{h_2})dv + \int_{-\infty}^{y-\delta h_2} f(v|x)dv \end{aligned}$$

From

$$\begin{aligned} \int_{-\delta}^{\delta} \Omega(t)dt &= \delta \\ \int_{-\delta}^{\delta} \Omega(t)W(t)dt &= 1/2 \\ 2\int_{-\delta}^{\delta} t\Omega(t)dt &= \delta^2 - \mu_2(W) \end{aligned}$$

Thus

$$\begin{aligned}
\sigma^2(x,y) &= E\{\Omega^2(\frac{y-Y}{h_2})|X=x\} - \{E\Omega(\frac{y-Y}{h_2})|X=x\}\}^2 \\
= F(y|x)(1-F(y|x)) &- h_2 F^{01}(y|x)\int \Omega(t)(1-\Omega(t))dt + O(h_2^2) \\
&= F(y|x)(1-F(y|x)) + O(h_2)
\end{aligned}$$

Lemma 3

Let $\hat{f}_{h_1,h_2}(y|x) = \hat{m}^{01}_{h_1,h_2}(y|x)$; in fact, $\hat{f}_{h_1,h_2}(y|x)$ is the local linear kernel estimator of the conditional density discussed by Fan, Yao and Tong (1996). Also, let $q_p^*(x)$ be some random point between $\tilde{q}_p(x)$ and $q_p(x)$, then under the conditions of Theorems 2.1 and 2.2, we have respectively

$$\hat{f}_{h_1,h_2}(q_p^*(x)|x) = f(q_p(x)|x) + o_p(1)$$

and

$$\hat{f}_{h_1,h_2}(q_p^*(0+)|0+) = f(q_p(0+)|0+) + o_p(1).$$

Proof. This follows Lemma 6 of Samanta (1989) and equations (6.4) and (6.5) of Fan (1993).

Proof of Theorem 2.1 (Similarly for Theorem 2.2)

From Lemmas 1 and 2, we have

$$\begin{aligned}
E\{\tilde{F}_{h_1,h_2}(y|x) - F(y|x)\}^2 &= \{1/2\{F^{20}(y|x)\}\mu_2(K)h_1^2 + 1/2\{F^{02}(y|x)\}\mu_2(W)h_2^2\}^2 \\
&+ \frac{R(K)}{nh_1 g(x)}\Big(F(y|x)(1-F(y|x)) - f(y|x)\alpha(W)h_2\Big) \\
&+ o(h_2^2/nh_1 + h_1^4 + h_2^4)
\end{aligned}$$

Then Theorem 2.1 follows from Lemma 3 and the following equation

$$\tilde{q}_p(x) - q_p(x) \simeq -\frac{\{\hat{m}_{h_1,h_2}(q_p(x)|x) - F(q_p(x)|x)\}}{f(q_p^*(x)|x)}.$$

This completes the proofs of Theorem 2.1 and 2.2.

For $h_1 >> h_2$, the leading terms in the MSE of $\tilde{q_p}(x)$ in Theorem 2.1 are essentially the squared bias and variance as for the check function approach, and they involve h_1 only. In fact, these differ from the terms in (2.5) in that the bias term of Theorem 2.1, the quantity $F^{20}(q_p(x)|x)/f(q_p(x)|x)$, involves the second derivative with respect to x of the conditional distribution function, instead of $q_p''(x)$ (second derivative with respect to x of the quantile function itself). However, Theorem 2.1 gives some guidance on how to choose h_2.

2.3.4 Bandwidth Selection

The rules–of–thumb developed in Section 2.2.4 may be used to select h_1, all that changes is that $F^{20}(q_p(x)|x)\ /f(q_p(x)|x)$ replaces $q_p''(x)$ in (2.6) and (2.7) and thus the arguments remain valid (or otherwise) for this function as they did for the previous.

That said, what we have glossed over so far is that one version of optimality, convergence rate–wise, is to choose h_2 as large as possible (subject to $h_2 \to 0$ when $n \to \infty$) provided the second term in Theorem 2.1 is positive and not inflating the bias term. This implies $h_2 \sim h_1$. In particular $h_2 = h_1$ results in replacing $F^{20}(q_p(x)|x)/f(q_p(x)|x)$ in the optimal h_1 formula by $\{F^{20}(q_p(x)|x) + F^{02}(q_p(x)|x)\}/f(q_p(x)|x)$.

Assuming $h_2 < h_1$, an alternative argument of rule–of–thumb for h_2 is pursued which continues to give MSE of $O(n^{-4/5})$. Intuitively, choice of h_2 should not be very critical since it concerns a smoothing at the distribution function level; moreover, various formulas for h_2 have been tried in practice and we have observed

little impact. It still remains to specify a value for h_2 and to ensure that the chosen value is never so extreme as to make an unnecessary impact. The practical experience and numerical computation show that h_2 should not be specified too small relative to h_1.

We thus concentrate on the order $h_1^2 h_2^2 + h_2/(nh_1)$ terms of $\mathrm{MSE}(\tilde{q}_p(x))$. While the latter is always negative, the former can be negative or positive at different xs. The optimal h_2 is either $h_2 = \infty$ or

$$h_2 = \left(\frac{R(K)\alpha(W)}{\mu_2(K)\mu_2(W)}\right) \frac{f(q_p(x)|x)}{g(x)|F^{20}(q_p(x)|x)F^{02}(q_p(x)|x)|} \frac{1}{nh_1^3} \tag{2.11}$$

respectively.

Regarding h_1 and h_2 as functions of p, the latter gives

$$\frac{h_{2,p_1} h_{1,p1}^3}{h_{2,p_2} h_{1,p2}^3} = \frac{f(q_{p_1}(x)|x)|F^{20}(q_{p_2}(x)|x)F^{02}(q_{p_2}(x)|x)|}{f(q_{p_2}(x)|x)|F^{20}(q_{p_1}(x)|x)F^{02}(q_{p_1}(x)|x)|}. \tag{2.12}$$

As in automatic selection of h_1 (Section 2.2.4) suppose that $|F^{20}(q_{p_1}(x)|x)| = |F^{20}(q_{p_2}(x)|x)|$ and also use a normal approximation $\sigma_x^{-1}\phi(\Phi^{-1}(p))$ to $f(q_p(x)|x)$. We likewise require a parametric approximation to $|F^{02}(q_p(x)|x)|$ and take the double exponential distribution as a guideline rather than the normal since the latter has a zero derivative at its median i.e. $|F^{02}(q_p(x)|x)| = \frac{1}{\lambda^2}\{(1-p)I(p \geq 1/2) + pI(p < 1/2)\}$. Then (2.12) reduces to

$$\frac{h_{2,p} h_{1,p}^3}{h_{2,1/2} h_{1,1/2}^3} = \frac{\sqrt{2\pi}\phi(\Phi^{-1}(p))}{2\{(1-p)I(p \geq 1/2) + pI(p < 1/2)\}}.$$

Take $h_{1,1/2} = h_{2,1/2}$. This somewhat arbitrary choice can be justified in various ways.

1). Estimating the regression median is rather akin to estimating the regression mean, and in the latter case h_2 is immaterial.

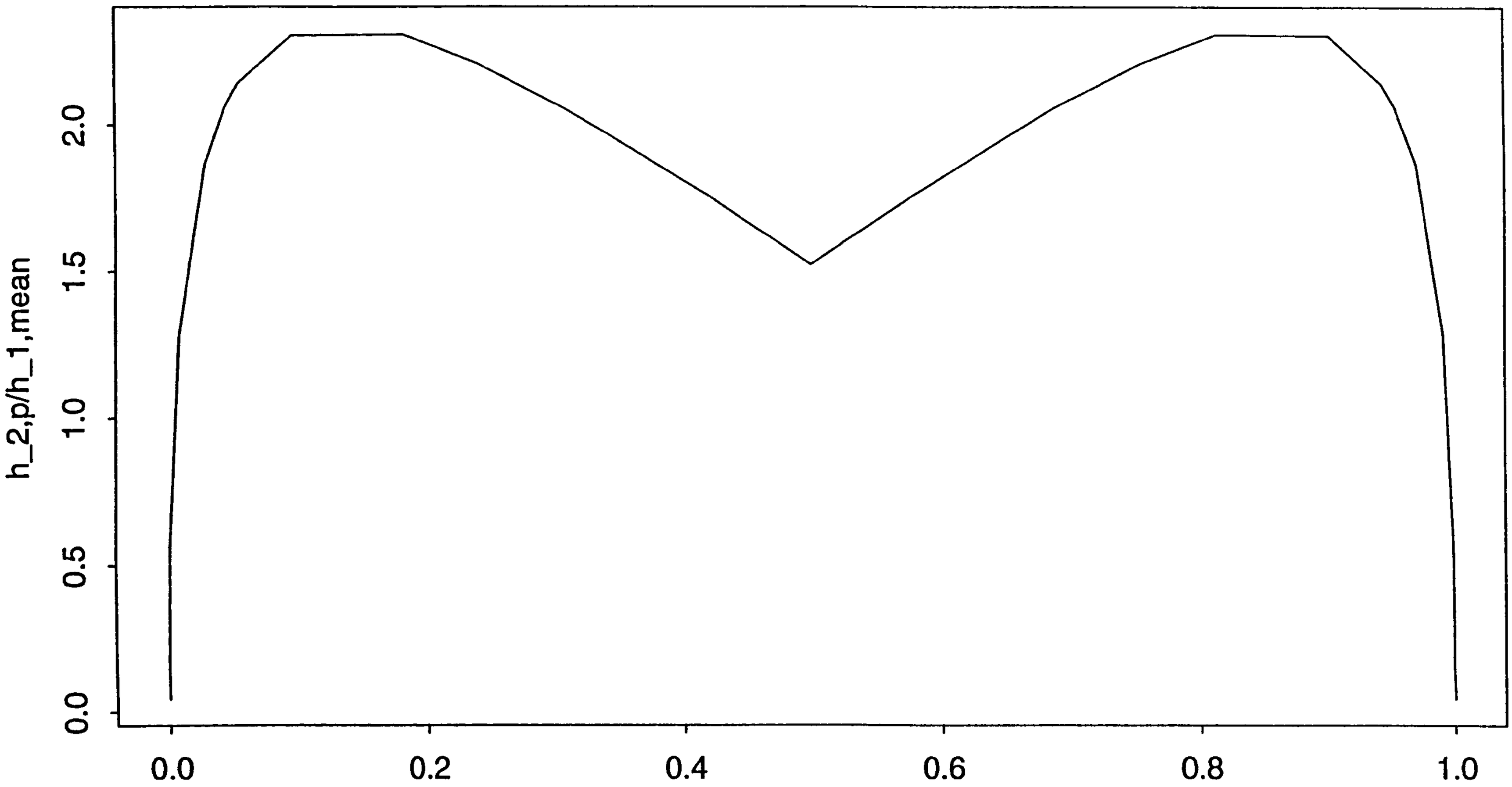


Figure 2.2: Relationship between h_2,p/h_1,mean against p by rule-of-thumb

2). h_2 should not be arbitrarily large for extreme quantiles from the asymptotic variance of $\tilde{q}_p(x)$.

3). For $p = 1/2$, $\Phi^{02}(q_{1/2}(x)|x) = 0$ and hence setting $h_{2,1/2} = h_{1,1/2}$ makes no difference to the bias.

The resulting rule–of–thumb recommended is

$$h_{2,p} = \frac{\sqrt{2\pi}\phi(\Phi^{-1}(p))}{2\{(1-p)I(p \geq 1/2) + pI(p < 1/2)\}} \frac{h_{1,1/2}^4}{h_{1,p}^3}. \tag{2.13}$$

which gives a "reasonable" formula for h_2. Inserting the h_p in Section 2.2.4 as $h_{1,p}$ of (2.13), the relationship between h_2 and p is shown in Figure 2.2, and interestingly, $h_{2,p} \to 0$ as $p \to 0, 1$, but when $p = 0.97$, $h_{2,p} = 1.236 h_{1,1/2}$.

2.4 Numerical Examples

To illustrate the methodology of the previous sections the five sets of data taken from the literature discussed in Section 1.4 are used. For each set several quantile estimates are calculated for $\{p : 0.5,\ (0.75, 0.25),\ (0.9, 0.1),\ (0.95, 0.05)$ or $(0.97, 0.03)\}$.

The standard normal kernel is used as K and the uniform kernel as W (where necessary) in all computations. Two S programs were written respectively for the two methods of iterative computation of quantiles, and two further S programs were used for the computation of the local linear fitted regression mean and its derivative.

2.4.1 Bandwidths

The following are the h_1 bandwidths chosen by our selection method for each dataset; values of h_2 follow from these by use of (2.13).

Set 1: the triceps skinfold data, $h_{mean} = 2.5$, $h_{0.5} = 2.7$, $h_{0.75} = h_{0.25} = 2.8$, $h_{0.9} = h_{0.1} = 3.1$ and $h_{0.97} = h_{0.03} = 3.6$. When $h_{2,0.5} = 2.7$, $h_{2,0.75} = h_{2,0.25} = 3.88$, $h_{2,0.9} = h_{2,0.1} = 3.95$, and $h_{2,0.97} = h_{2,0.03} = 3.16$.

Set 2: the immunoglobin data, $h_{mean} = 0.5$, $h_{0.5} = 0.54$, $h_{0.75} = h_{0.25} = 0.56$, $h_{0.9} = h_{0.1} = 0.62$ and $h_{0.95} = h_{0.05} = 0.67$.

Set 3: the body weight data, $h_{mean} = 1.8$, $h_{0.5} = 1.97$, $h_{0.75} = h_{0.25} = 2$, $h_{0.9} = h_{0.1} = 2.23$ and $h_{0.97} = h_{0.03} = 2.6$.

Set 4: the heart transplant data, $h_{mean} = 6$, $h_{0.5} = 6.57$, $h_{0.75} = h_{0.25} = 6.78$, $h_{0.9} = h_{0.1} = 7.44$ and $h_{0.95} = h_{0.05} = 8$.

Set 5: the motorcycle data, $h_{mean} = 1.2$, $h_{0.5} = 1.3$, $h_{0.75} = h_{0.25} = 1.35$, $h_{0.9} = h_{0.1} = 1.49$ and $h_{0.95} = h_{0.05} = 1.6$.

2.4.2 Discussion of the Results and Conclusion

The graphs of check function and double kernel fits are shown in Figures 2.3-2.7. The first impression from these figures is that the messages yielded by the check function and double kernel approaches about the conditional quantiles are broadly similar. That said, there is a distinct tendency for the double kernel smooths to be smoother than the check function. It seems that the kernel in

the vertical direction is beneficial at the very least in producing more pleasing pictures, an impression particularly gained from Figures 2.3 through 2.5 for data sets 1-3.

Comparison of this Figure 2.3 with Figure 2.2 of Cole & Green (1992) shows a very considerable degree of similarity in the results of these methods and theirs (which is an interesting semiparametric approach involving penalty functions, see Chapter 5). The interesting comparison of Figure 2.4 with Royston & Altman's (1994) Figure 2.5 — where the regression mean is estimated by parametric methods — is that all our quantiles including the median display a marked peak at the larger ages which is not apparent in Royston & Altman's models. Comparing it with the results of Isaacs et al (1983), the main difference lies near the right–hand boundary where their results seem to be a little oversmooth and hence too flat.

Figures 2.6(a) and (b) are broadly similar and in line with Figure 2.5 of Cole and Green (1992). It was an earlier version of this example that led us to prefer the current rule for choosing h_2 to other suggestions. Earlier versions resulted in larger h_2 for extreme p, and this resulted, as observed quite generally, in a considerable widening of the extreme quantiles particularly towards the left–hand end (Figure 2.6(b)). The current version results in $h_2 \to 0$ for extreme p, and hence much more acceptable narrow extremes.

There is also a greater spread apparent in the lower spread localities of the motorcycle data in Figure 2.7(b) than in Figure 2.7(a). Referring to the scatterplot, it is difficult to pronounce too strongly which is the more realistic, but we have a slight preference for the "more conservative" bands of Figure 2.7(b). As Koenker, Ng & Portnoy (1992) point out, this is a prime candidate for the use of different window widths at different places to achieve more uniformly smooth curves: the

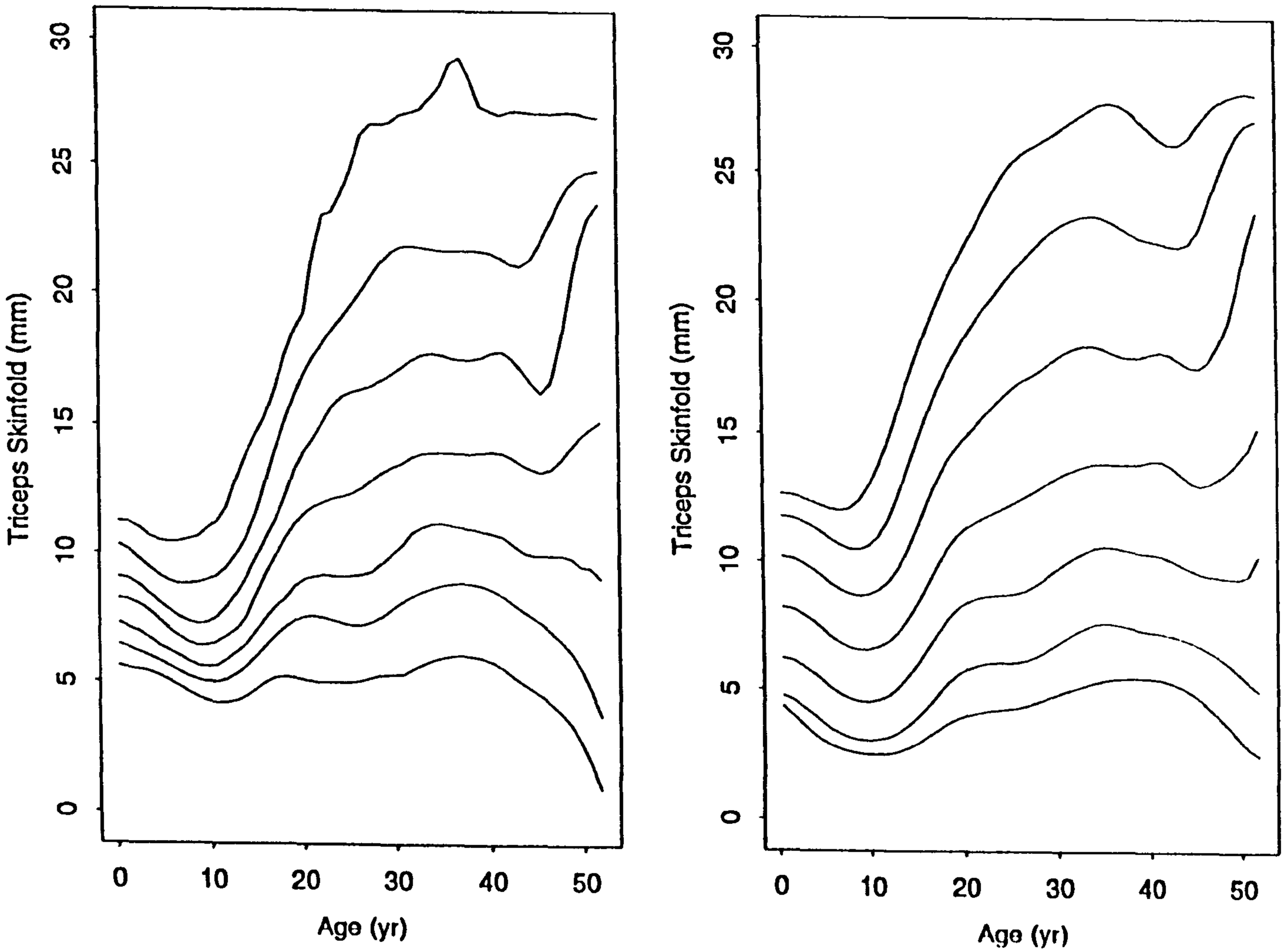


Figure 2.3: Smoothed reference centiles curves for triceps skinfold data at 3rd, 10th, 25th, 50th, 75th 90th and 97th percentiles

(a) Single-kernel smoothing (b) Double-kernel smoothing

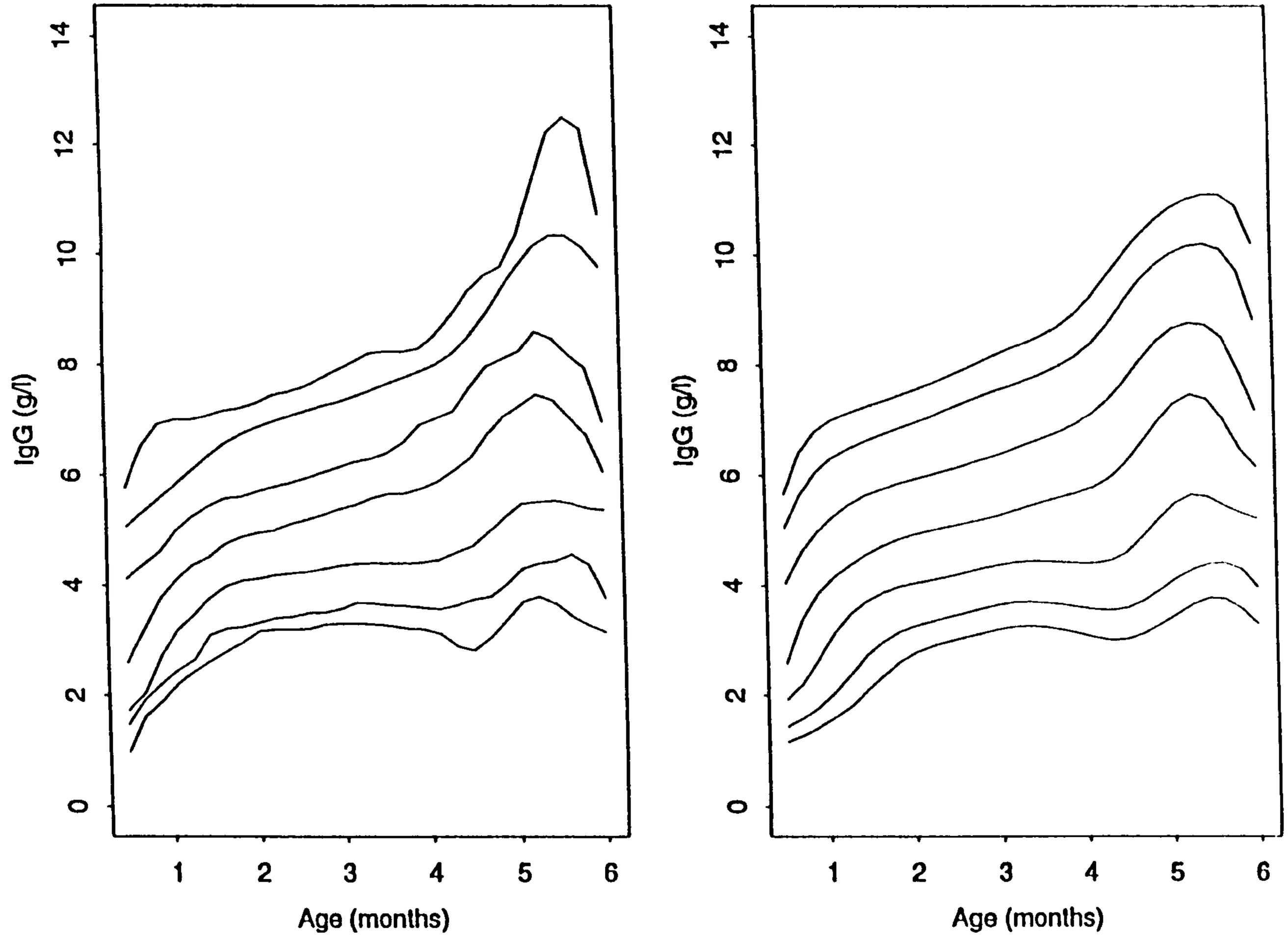
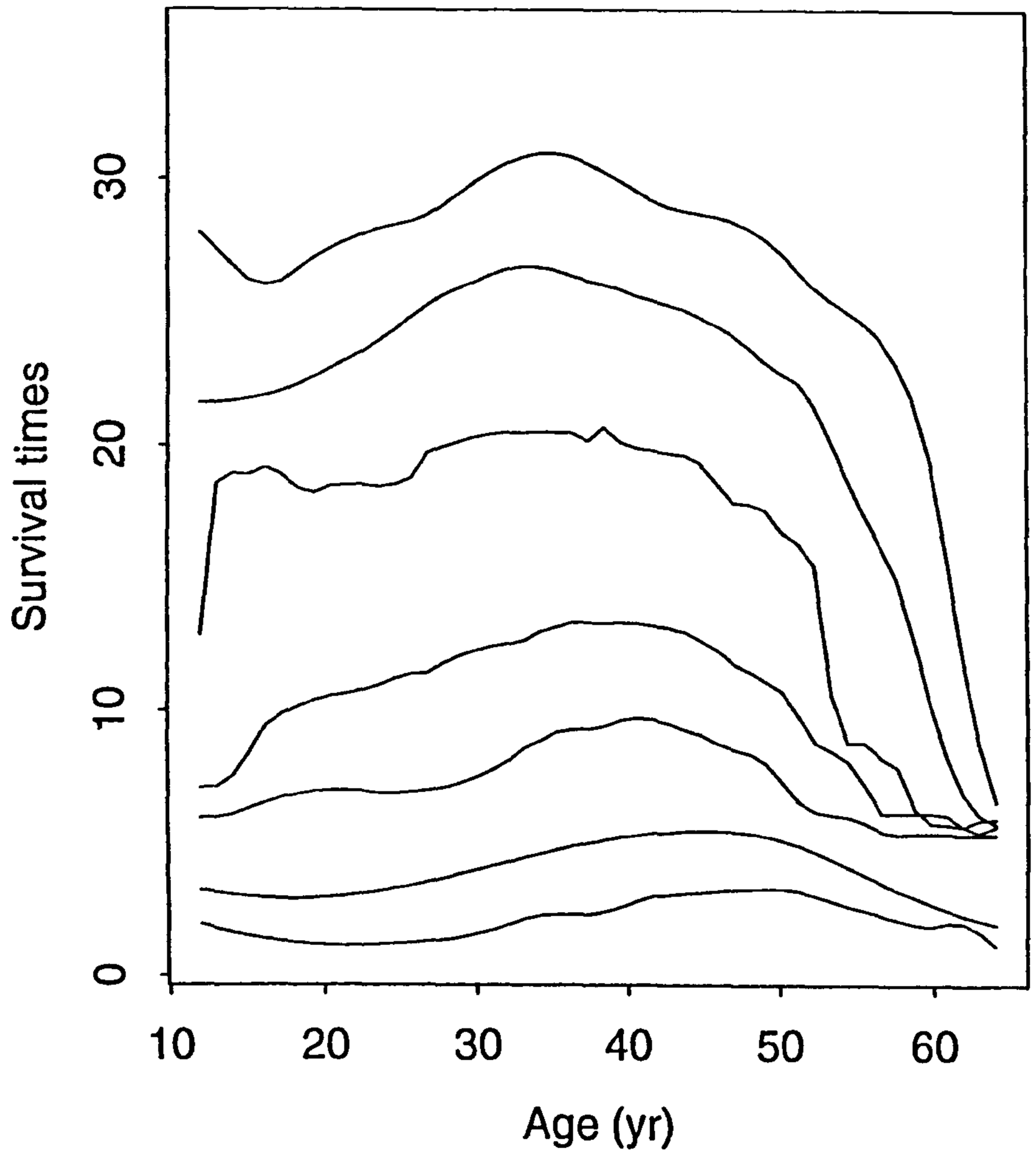


Figure 2.4: Smoothed reference centiles curves for immunoglobin-G data at 5th, 10th, 25th, 50th, 75th 90th and 95th percentiles

(a) Single-kernel smoothing (b) Double-kernel smoothing



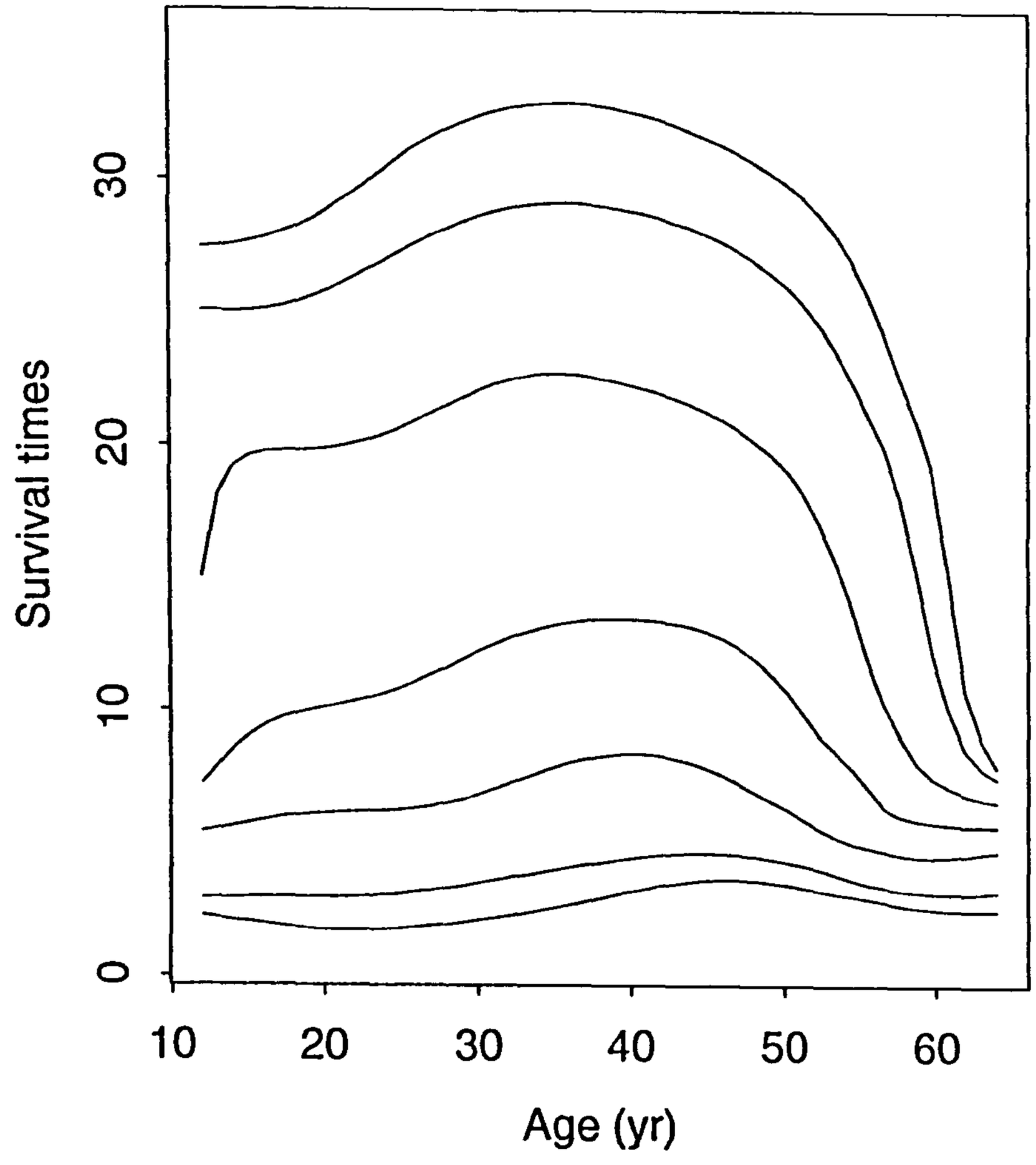


Figure 2.5: Smoothed reference centiles curves for heart transplant data at 5th, 10th, 25th, 50th, 75th 90th and 95th percentiles

(a) Single-kernel smoothing (b) Double-kernel smoothing

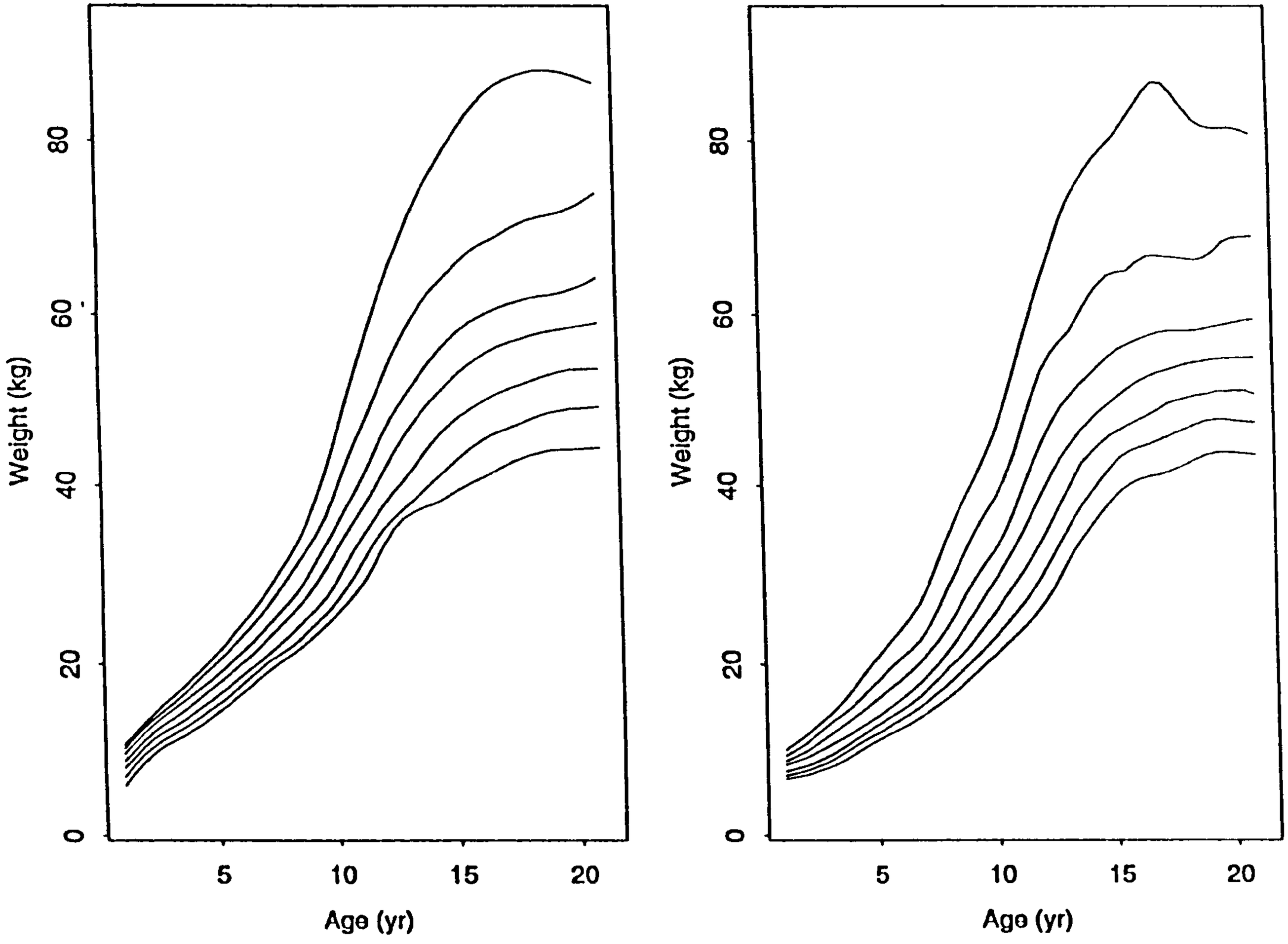


Figure 2.6: Smoothed reference centiles curves for body weight data at 3rd, 10th, 25th, 50th, 75th 90th and 97th percentiles

(a) Single-kernel smoothing (b) Double-kernel smoothing

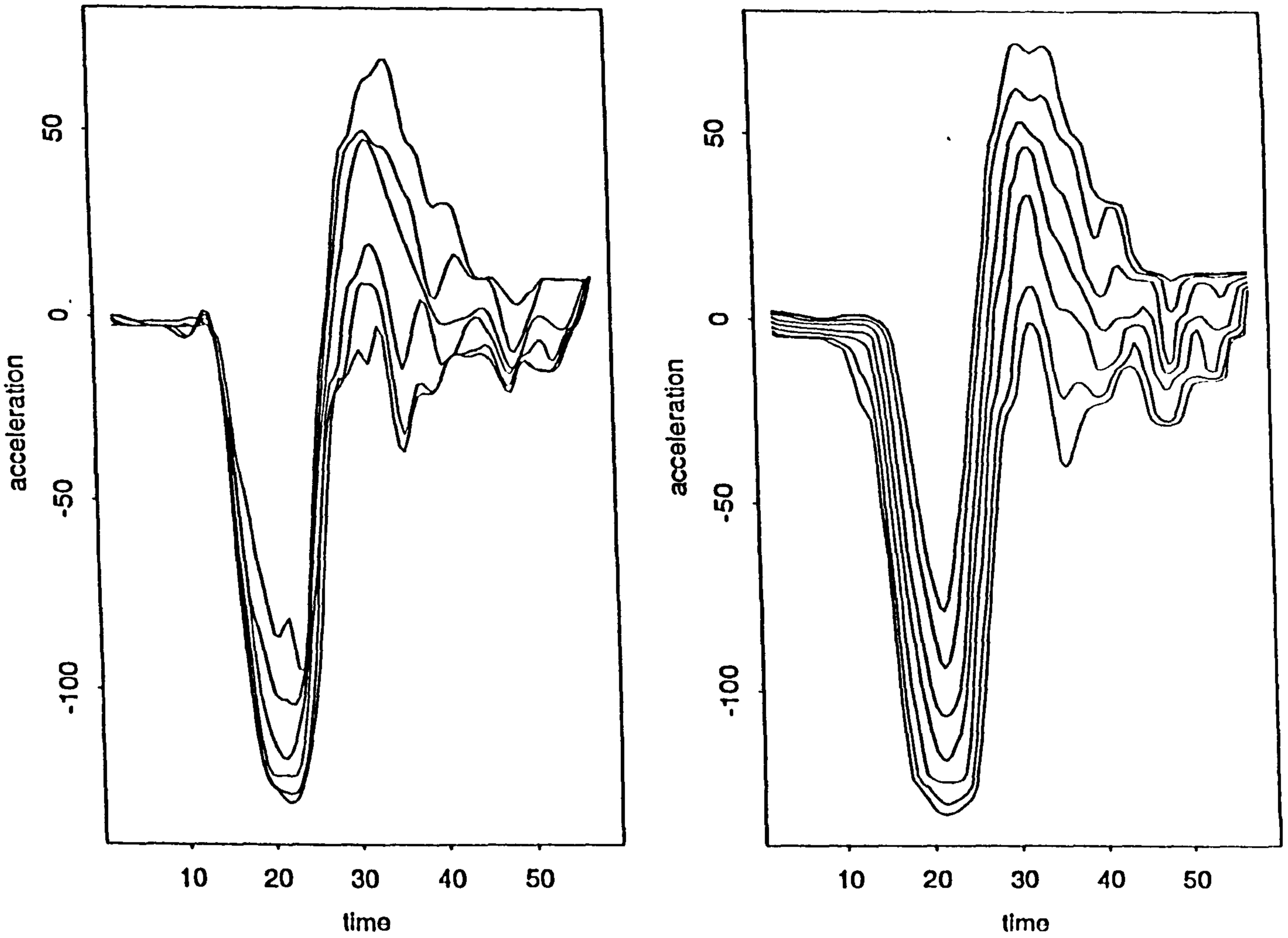


Figure 2.7: Smoothed reference centiles curves for motorcycle data at 5th, 10th, 25th, 50th, 75th 90th and 95th percentiles

(a) Single-kernel smoothing (b) Double-kernel smoothing

trough around time 20 and the flat portion to the right-hand side would clearly benefit from different degrees of smoothing.

The conclusions that can be drawn based on the results from data analysis are that the local linear conditional quantile estimation is feasible and practical, and the results are at the least comparable with those produced by other approaches. Besides, the way these methods work is interpretable and natural, asymptotic properties are tractable and informative, and computational algorithms are sensible.

While the approach is not novel, the implementation is provides rules for the selection of bandwidths which appear to give reasonable results. The differences between the two approaches developed here are interesting. In particular, the vertical smoothing (double kernel approach) appears to be advantageous in providing reasonable extra smoothing over that inherent in the check function approach. It also appears to be advantageous vis–a–vis non–crossing of quantiles. The double kernel method is a practically useful method, and one worthy of further fine–tuning for even better performance in the future.

Chapter 3

A Comparison of Local Constant and Local Linear Regression Quantile Estimators

3.1 Introduction

To investigate the difference between local linear smoothing approach and local constant fit, suppose that $\{(X_i, Y_i)\}_1^n$ are i.i.d. observations and $q_p(x)$ is the p–quantile of the conditional distribution $F(y|x)$ of Y given X. Recall the check function $\rho_p(t) = \{|t| + (2p-1)t\}/2$ suggested by Koenker & Bassett (1978) and define a kernel estimator of $q_p(x)$ by

$$\overline{q_p(x)} = \text{argmin}_a \sum_1^n \rho_p(Y_i - a) K(h^{-1}(x - X_i)). \tag{3.1}$$

Here, K is a kernel function with bandwidth h, and let K be a symmetric unimodal density function. The estimator $\hat{q}_p(x)$ in (3.1) is in fact a *local constant*

version of conditional quantile estimation when the constant a is fitted locally. The local linear version of this was already given at (2.4) and is given by $\hat{q}_p(x) = \hat{a}$ such that

$$(\hat{a}, \hat{b}) = \text{argmin}_{a,b} \sum_{1}^{n} \rho_p(Y_i - a - b(X_i - x))K(h^{-1}(x - X_i)). \qquad (3.2)$$

If $\rho_p(z)$ is replaced by z^2, then the well known local constant (Nadaraya–Watson) and local linear regression *mean* estimators would ensue. The advantages of local linear over local constant fitting have been discussed by many authors amongst them Fan (1992), Ruppert & Wand (1994) and Cleveland & Loader (1995). One naturally wonders whether the two approaches to smoothing conditional quantiles defined above afford analogous results. A thorough investigation theoretically and in practice shows that the general answer, unsurprisingly, is 'yes', and points out the advantages of local linear over local constant fitting for the mean and quantiles.

However, unlike one's imagination, the difference of two fittings, particularly practical difference, is very small in our experience, so one can not expect much improvement by using $\hat{q}_p(x)$ instead of $\overline{q}_p(x)$ except in boundaries.

The mean squared errors (MSEs) of the two methods for the interior of the design space are derived in Sections 3.2.1 and 3.2.2 while the boundary effects are discussed in Section 3.2.3. The two methods are compared in Section 3.3 using simulated data and the data described in Section 1.4. Regression means are also considered in Section 3.3.2. A shortened version of this chapter forms Yu & Jones (1997b).

3.2 Theoretical Comparison

The asymptotic MSEs of $\overline{q}_p(x)$ and $\hat{q}_p(x)$ are compared as $n \to \infty$, $h = h(n) \to 0$ and $nh \to \infty$. The investigation centres on the behaviour at interior and boundary points of the design space. Three out of four of these cases already appear in the literature, but bringing them together and focussing on their comparison is novel. Indeed, MSEs for $\hat{q}_p(x)$ were presented in Section 2.2.3.

The underlying assumptions are the usual standard ones such as continuous derivatives of quantities involved (Fan, Hu & Truong, 1994). The marginal density function $g(x)$ of the X is assumed to be continuous with support $[0, 1]$, and let the support of K be $[-1, 1]$. Define the interior to be $h < x < 1 - h$, the case in which x being an interior point away from 0 and 1 is investigated in the following two sections.

3.2.1 Asymptotic MSE: x in Interior

The asymptotic biases and variances of $\overline{q}_p(x)$ and $\hat{q}_p(x)$ are given in Table 3.1. Results for $\overline{q}_p(x)$ are taken from Jones & Hall (1990) and for $\hat{q}_p(x)$ from Fan, Hu & Truong (1994) and Section 2.2.3. The MSE of any estimator (*) may be obtained, of course, from $MSE(*) = Bias(*)^2 + Variance(*)$.

Estimator	bias	variance
local constant fit	$\frac{1}{2}h^2\mu_2(K)\left\{-\frac{F^{20}(q_p(x)\|x)}{f(q_p(x)\|x)} + 2\frac{g'(x)q_p'(x)}{g(x)}\right\}$	$\frac{p(1-p)R(K)}{nhg(x)f^2(q_p(x)\|x)}$
local linear fit	$\frac{1}{2}h^2\mu_2(K)q_p''(x)$	$\frac{p(1-p)R(K)}{nhg(x)f^2(q_p(x)\|x)}$

Table 3.1: Pointwise Bias and Variance of $\overline{q}_p(x)$ and $\hat{q}_p$

Note that the two asymptotic variances are identical and the differences in the mean squared errors between local constant and linear fits depend only on their respective biases. Had the term $F^{20}(q_p(x)|x)/f(q_p(x)|x)$ instead been $-q_p''(x)$, these biases would have followed precisely the form of biases in the regression mean case (e.g. Fan, 1992) where q_p is replaced by mean m. An unappealing property of the local constant bias is the second term, not present in the local linear case, which depends on the marginal density g. Clearly, this term can make a lot of difference, perhaps detrimentally, if the design density is highly clustered and the conditional quantile is increasing or decreasing dramatically. As expected, $q_p''(x)$ is associated with local linear case, but $F^{20}(q_p(x)|x)/f(q_p(x)|x)$ is, perhaps, more naturally associated with methods used to estimate the conditional quantiles, usually estimating the conditional distribution function and then inverting (see Chapter 2).

Writing $b(x)$ for either $q_p''(x)$ or $-\frac{F^{20}(q_p(x)|x)}{f(q_p(x)|x)} + 2\frac{g'(x)q_p'(x)}{g(x)}$, the best possible MSE in either case is given by

$$5/4 \left\{ \frac{p(1-p)R(K)}{f^2(q_p(x)|x)g(x)n} \right\}^{4/5} \{\mu_2(K)b(x)\}^{2/5} n^{-4/5}$$

achieved when the optimal h is

$$h^5 = \frac{p(1-p)R(K)}{\mu_2(K)^2 b^2(x) f^2(q_p(x)|x)g(x)n}.$$

3.2.2 Further Comparison of Leading Bias Terms

A uniform design is the obviously most practical and important case for which $g'(x)q_p'(x) = 0$. As noted earlier, the two bias terms still differ, but they are by no means unrelated; in fact, in general, the first term between the curly brackets

of the bias term for local constant fitting in Table 3.1 may be written as

$$-\frac{F^{20}(q_p(x)|x)}{f(q_p(x)|x)} = q_p''(x) + \{q_p'(x)\}^2\{\log f(q_p(x)|x)\}^{01} + 2q_p'(x)\{log f(q_p(x)|x)\}^{10} \tag{3.3}$$

(Jones & Hall, 1990).

Two interesting cases arise for which the biases of the two fittings are equal i.e. $q_p'(x)g'(x) = 0$, and the last terms in (3.3) vanish when

(i) The conditional density function $f(q_p(x))$ is uniform, or

(ii) $q_p(x)$ is flat or it has a maximum, minimum or inflection point at x.

In the rest of this section assume that the regression model for which the conditional distribution of Y given $X = x$ is of the same form for each x except for a differing location (i.e. $Y = m(x) + \epsilon$ and ϵ has a density $\chi(t)$, say), then (3.3) simplifies further to

$$-\frac{F^{20}(q_p(x)|x)}{f(q_p(x)|x)} = q_p''(x) - \{q_p'(x)\}^2\{\log f(q_p(x)|x)\}^{01}. \tag{3.4}$$

This yields the following conclusions (when $g'(x)q_p'(x) = 0$):

(1) The two smoothings have equal asymptotic bias, variance and MSE when $\chi(u)$ is a uniform.

(2) If $|q_p''(x)| >> |q_p'(x)|$, the two smoothing approaches have approximately equal asymptotic bias, variance and MSE.

(3) When $q_p''(x) = 0$, $MSEL(x) \leq MSEC(x)$.

(4) More generally, when $q_p''(x)log^{01}\{f(q_p(x)|x)\} < 0$,

$$|biasL(x)| \leq |biasC(x)| \text{ and } MSEL(x) \leq MSEC(x).$$

Here $biasC$ $(MSEC)$ and $biasL$ $(MSEL)$ are written for the biases (mean squared errors) of the local constant and local linear fits, respectively. Note that it is possible for $MSEL(x) \leq MSEC(x)$ when $q_p''(x) log^{01}\{f(q_p(x)|x)\} \geq 0$, depending on the relative sizes of $q_p''(x)$ and $\{q_p'(x)\}^2 log^{01}\{f(q_p(x)|x)\}$.

Distribution of ϵ	$log^{01}\{f(q_p(x)\|x)\}$
Normal (0,1)	$-\Phi^{-1}(p)$
Uniform	0
Cauchy	$-2\frac{\tan((p-1/2)\pi)}{1+\tan^2((p-1/2)\pi)}$
Laplace	$\lambda\{I(p<1/2) - I(p>1/2)\}$
Exponential (λ)	$-\lambda$
Lognormal (0,1)	$-\frac{1+\Phi^{-1}(p)}{e^{\Phi^{-1}(p)}}$

Table 3.2: $log^{01}\{f(q_p(x)|x)\}$ for Error Distributions

The size of $log^{01}\{f(q_p(x)|x)\}$ is evaluated for the following categories of distributions in Table 3.2.

(i). Symmetric Short-Tailed: Normal, Uniform.

(ii). Symmetric Long-Tailed: Cauchy, Laplace

(iii). Skewed: Exponential, Lognormal.

In Table 3.2, $\Phi^{-1}(p)$ is the p–quantile of N(0,1).

It follows that for the symmetric distributions considered (indeed for all symmetric unimodal distributions) $log^{01}\{f(q_p(x)|x)\} \leq 0$ (> 0) when $p \geq 1/2$ $(< 1/2)$. The region in which $MSEL(x) \leq MSEC(x)$ is *guaranteed* is then when $p \geq 1/2$ $(< 1/2)$ and $q_p''(x) \geq 0$ (< 0). The $MSEL(x)$ will be smaller than

$MSEC(x)$ for other combinations of q'' and p. For the lognormal distribution, the above holds when threshold for p is replaced by $\Phi^{-1}(-1) = 0.1587$, and for the exponential by $p = 0$.

Consider the model $Y_i = m(X_i) + \epsilon_i$ where $\epsilon_i \sim N(0,1)$ and $m(x) = 1 - x + \exp(-200(x-1/2)^2)$ (Härdle, 1990). The pth quantile is given by $q_p(x) = m(x) + \Phi^{-1}(p)$ and

$$q_p''(x) = 400\exp(-200(x-1/2)^2)(400(x-1/2)^2 - 1)$$

while $log^{01}\{f(q_p(x)|x)\} = -\Phi^{-1}(p)$. The region for which $MSEL(x)$ is guaranteed to be less than or equal than $MSEC(x)$ depends on the sign of $q_p''(x)log^{01}\{f(q_p(x))\}$ as in Table 3.3. However, these areas are conservative and in this case we can do

$q_p''(x)log^{01}\{f(q_p(x)\|x)\}$	$p < 0.5$	$p = 0.5$	$p > 0.5$
$x < 0.45$	> 0	0	< 0
$0.45 < x < 0.55$	< 0	0	> 0
$x > 0.55$	> 0	0	< 0

Table 3.3: Sign of $q_p''(x)log^{01}\{f(q_p(x)|x)\}$ for Härdle's Model

some further work to find out when $|q_p''(x) - \{q_p'(x)\}^2 log^{01}\{f(q_p(x)|x)\}| \geq |q_p''(x)|$ ($x \in [0,1]$). For $p > 0.5$, the above approximation is almost exact. $MSEL(x) \leq MSEC(x)$ except for x in the interval $(0.45, 0.55)$: numerical calculations shrink the interval of inferiority to around $(0.46, 0.54)$, these values depend on p but to a surprisingly limited extent (e.g. the lower boundary is 0.466 when $p = 0.97$ and is 0.459 when $p = 0.75$). When $p < 0.5$, similar calculations show that, approximately $MSEL(x) \leq MSEC(x)$ for $x \in (0.28, 0.55) \cup (0.72, 1]$. However, these calculations ignore the boundary effect which works in favour of linear fitting as shown in the following subsection.

3.2.3 Asymptotic MSE: x Near Boundary

The asymptotic properties of MSE in both approaches are studied for boundary points. Writing $x = ch$, $0 \leq c < 1$ (the other boundary at 1 could be accommodated in an analogous way), and set

$$a_l(c;K) = \int_{-1}^{c} u^l K(u)du.$$

Further let

$$\alpha_c(K) = \frac{a_2^2(c;K) - a_1(c;K)a_3(c;K)}{a_0(c;K)a_2(c;K) - a_1^2(c;K)} \text{ and } \beta_c(K) = \frac{\int_{-1}^{c}(a_2(c;K) - a_1(c;K)u)^2K^2(u)du}{\{a_0(c;K)a_2(c;K) - a_1^2(c;K)\}^2}.$$

The asymptotic bias and variance of $\hat{q}_p(x)$ at the boundary points are taken from Fan, Hu & Truong (1994) and Section 2.2.4, and the corresponding results for $\overline{q_p(x)}$ are derived by adapting the work of Jones and Hall (1990) as follows.

Write

$$H_p(a) = \sum_{i=1}^{n} W_i\psi_p(Y_i - a)$$

where $\psi_p(z) = pI_{(0,\infty)}(z) - (1-p)I_{(-\infty,0)}(z)$ and $W_i = (nh)^{-1}K(h^{-1}(x - X_i))$.

Then

$$\begin{aligned}
\mu_p(a) &\equiv E\{H_p(a)|X_1,...,X_n\} = \sum_{i=1}^{n} W_i\{p - F(a|X_i)\} \\
&\simeq -\sum_{i=1}^{n} W_i(x)\left\{(a - q_p(x))f(q_p(x)|x) + (X_i - x)F^{10}(q_p(x)|x)\right\} \\
&= -s_{n0}(x)(a - q_p(x))f(q_p(x|x) + s_{n1}(x)F^{10}(q_p(x)|x).
\end{aligned}$$

Here

$$s_{nl}(x) \equiv \sum_{i=1}^{n}(x - X_i)^l W_i(x) \simeq h^l a_l(c;K)g(x)$$

so that

$$\mu_p(a) \simeq -a_0(c;K)g(x)(a - q_p(x))f(q_p(x)|x) + ha_1(c;K)g(x)F^{10}(q_p(x)|x)$$

which when set equal to zero yields

$$a = q_p(x) + h(a_1(c;K)/a_0(c;K))F^{10}(q_p(x)|x)/f(q_p(x)|x).$$

For the variance, note that

$$\text{var}\{H_p(a)|X_1,...,X_n|\} \simeq \sum_{i=1}^{n} W_i^2(x)p(1-p)$$

and

$$\sum_{i=1}^{n} W_i^2(x)dx \simeq (nh)^{-1}g(x)\kappa.$$

It can be shown that $\text{var}[E\{H_p(a)|X_1,...,X_n\}]$ is of smaller order and on dividing the right–hand side by $a_0^2(c;K)g^2(x)f^2(q_p(x)|x)$ gives the required asymptotic variance.

Estimator	bias	variance
local constant fit $\overline{q}_p(x)$	$h\frac{a_1(c;K)}{a_0(c;K)}\frac{F^{10}(q_p(x)\|x)}{f(q_p(x)\|x)}$	$\frac{p(1-p)\kappa}{nha_0^2(c;K)g(x)f^2(q_p(x)\|x)}$
local linear fit $\hat{q}_p(x)$	$\frac{1}{2}h^2\alpha_c(K)q_p''(x)$	$\frac{p(1-p)\beta_c(K)}{nhg(x)f^2(q_p(x)\|x)}$

Table 3.4: Pointwise Bias and Variance of $\hat{q}_p$ and $\check{q}_p$ Near the Boundary at Zero

For comparison, these results are presented in Table 3.4. Clearly, the boundary bias of local constant fitting compares unfavourably with the bias of local linear fitting as these biases are of order h and h^2 respectively. The local linear quantile estimate $\hat{q}_p(x)$'s boundary MSE constants are the same as those for the local linear mean estimation problem (Fan & Gijbels, 1992).

3.3 Practical Comparison

The standard normal kernel is used throughout the study of practical performance to compare the local constant and local linear quantile fitting methods. Two

simulated examples are considered first, then further applications are carried using the data available in Section 1.4.

3.3.1 Estimated Quantiles

(a) Simulated Data:

One hundred data points are generated from each of two versions of Härdle's (1990) model $Y_i = m(X_i) + \epsilon_i$ where $m(x) = 1 - x + \exp(-200(x - 1/2)^2)$. True 50th and 75th quantiles are obtained, and their local linear and local constant estimators are calculated for two different design spaces and bandwidths. h is selected subjectively in each case. Particularly,

(i) $\epsilon_i \sim N(0, 1)$, $X_i \sim U[0, 1]$ and $h = 0.05$, and

(ii) $\epsilon_i \sim U(0, 1)$, $X_i \sim N[0, 1]$ and $h = 0.15$.

The results are displayed in Figures 3.1 (a), (b) respectively. It is observed from Figure 3.1 (a) that the estimators are rather similar in the interior, particularly for the median, and less so towards boundary. The differences are more pronounced for $75th$ quantile (upper quartile). When the design and residual densities are reversed, Figure 3.1 (b), it is noticed that local constant and local linear estimators are less close, perhaps this is due to the greater steepness of $q'_p(x)$ and the non–uniform design.

Further, a second simulation situation is $n = 500$ points from the model

$$Y = 2 + sin(2X) + 2exp(-16X^2) + \epsilon$$

where $X \sim N(0, 1)$ truncated to [-2, 2] and $\epsilon \sim N(0, 0.25^2)$, independently. The

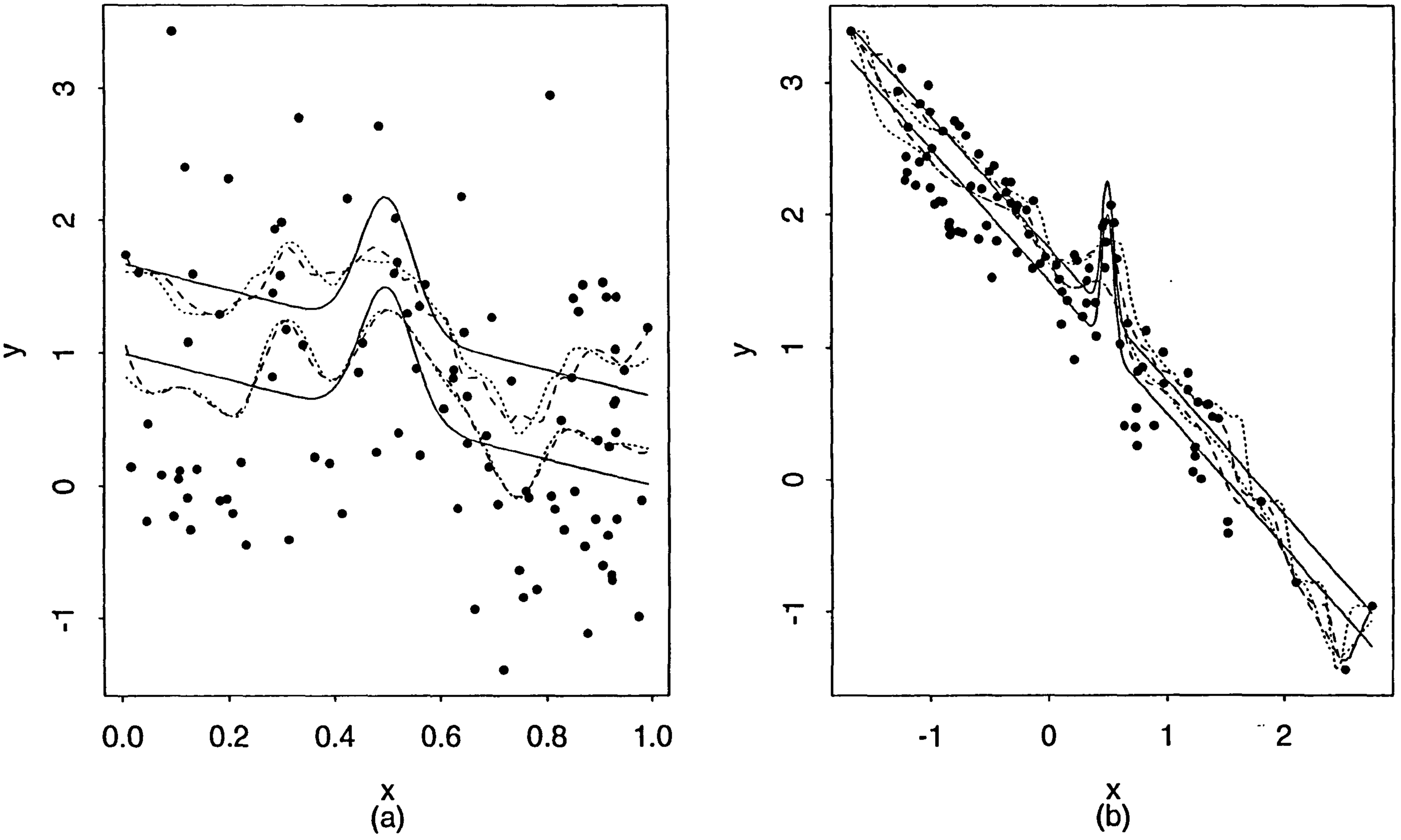


Figure 3.1 : Median and 75th percentile with $Y=1-X+\exp(-200(X-0.5)^2]+e$, n=100, underly curves (solid lines), local constant fit (dashed lines), local linear fit (dotted lines). (a): X~U(0,1), e~N(0,1), h=0.05. (b): X~N(0,1), e~U(0,1), h=0.15.

pointwise root MSE, averaged over 100 replications (with fixed Xs), for estimation of $q_{0.9}$ is shown in Figure 3.2. The bandwidths were chosen, separately for each estimator, by minimising the asymptotic integrated MSE for the given model. Clearly, differences between $\sqrt{MSE(\overline{q}_{0.9}(x))}$ and $\sqrt{MSE(\hat{q}_{0.9}(x))}$ are small in the interior, but can be substantial near the boundaries (here, towards the right-hand end).

(b) **Real Data:**

To compare the performance of local linear and constant fitting further the datasets in Section 1.4 are used with the same bandwidths obtained in Chapter 2. The estimators $\overline{q}_p(x)$ and $\hat{q}_p(x)$ are calculated for p in the symmetric class $\{5, 10, 25, 50, 75, 90, 95\}$ except for the first and fourth dataset the $3th$ and $97th$ replaces $5th$ and $95th$ quantiles respectively (these are the same quantiles used by other authors in similar studies), and the curves of these quantiles are plotted for each dataset superimposed on the scatter plot figures. As n is very large for sets and to avoid obscure the estimated curves the data points are not shown in Figures 3.3 or 3.6.

It is noticed that there is a considerable similarity between the quantiles estimated by the two methods except near the boundaries in some cases. In the interior, the differences would generally make no impact on qualitative conclusions and are, if anything, of smaller magnitude than the differences between the two versions of local linear fitting demonstrated in Chapter 2 . In Figure 3.6 the local linear fit is somewhat tighter than the local constant throughout much of the range of interest but neither are inconsistent with the data. At the boundaries especially, the differences can be considerable, in particular, for serum concentration data, Figure 3.4. A peak toward the right–hand edge (indicated by both local linear

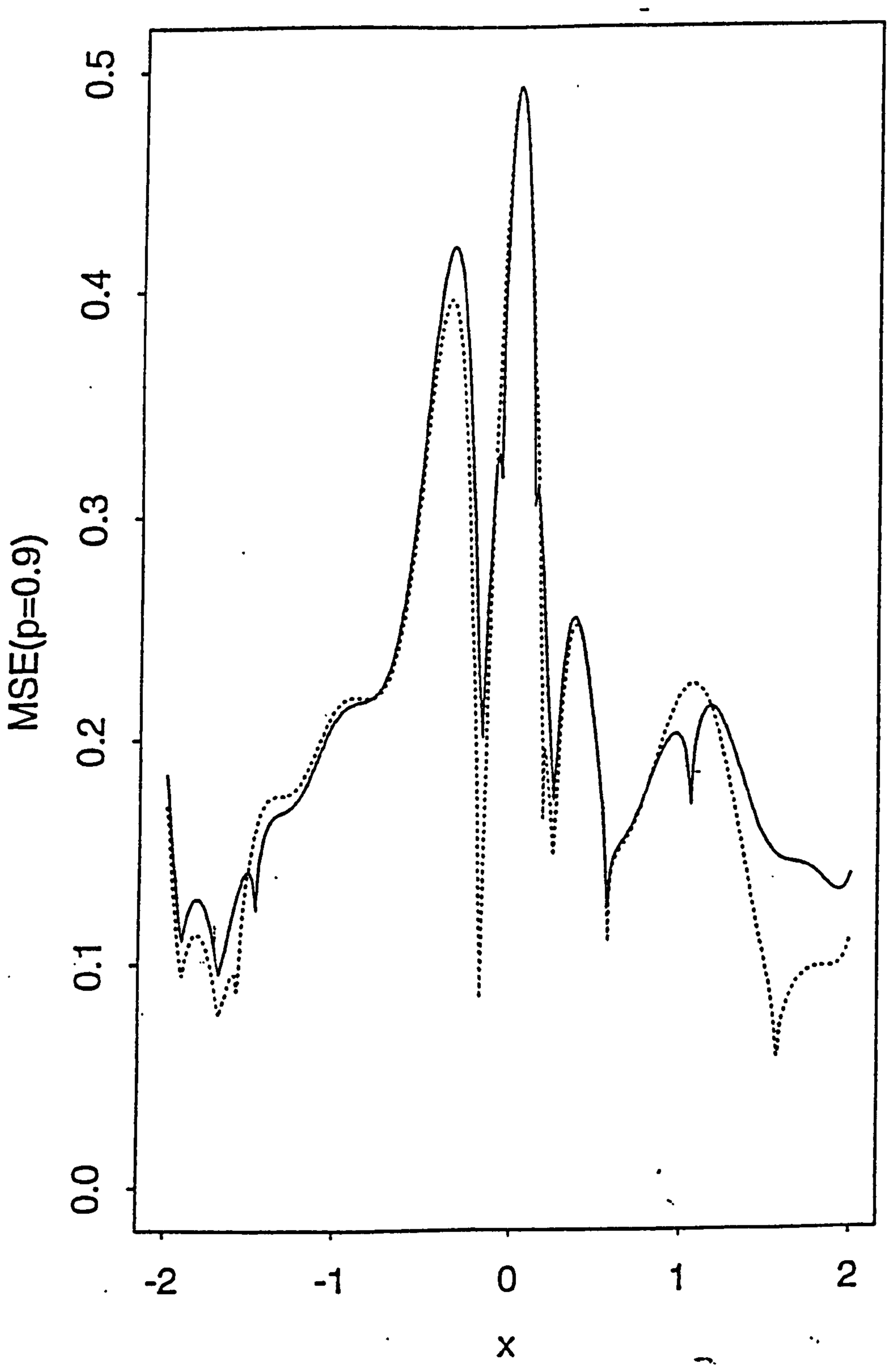


Figure 3.2: 100 simulated root MSEs for local constant fitting (solid lines) and local linear fitting (dotted lines) with n=500 for model $Y=2+\sin(2X)+2\exp(-16X^2)+e$, $X\sim N(0,1)$ truncated to [-2,2] and $e\sim N(0,0.25)$

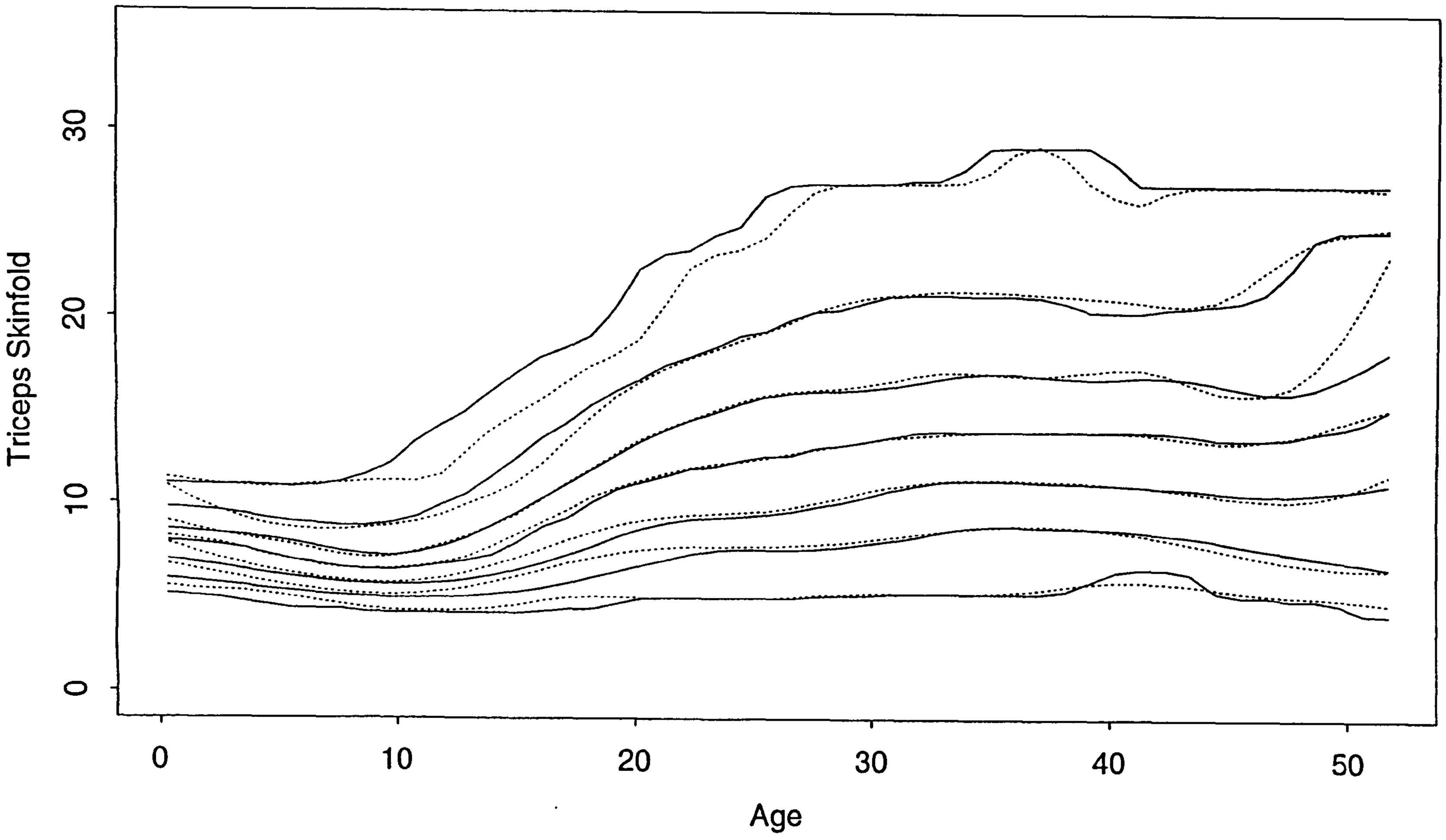


Figure 3.3: Seven smoothed quantile curves for triceps skinfold data by local constant fiting (solid lines) and local linear fitting (dotted lines) 3rd, 10th, 25th, 50th, 75th 90th and 97th percentiles

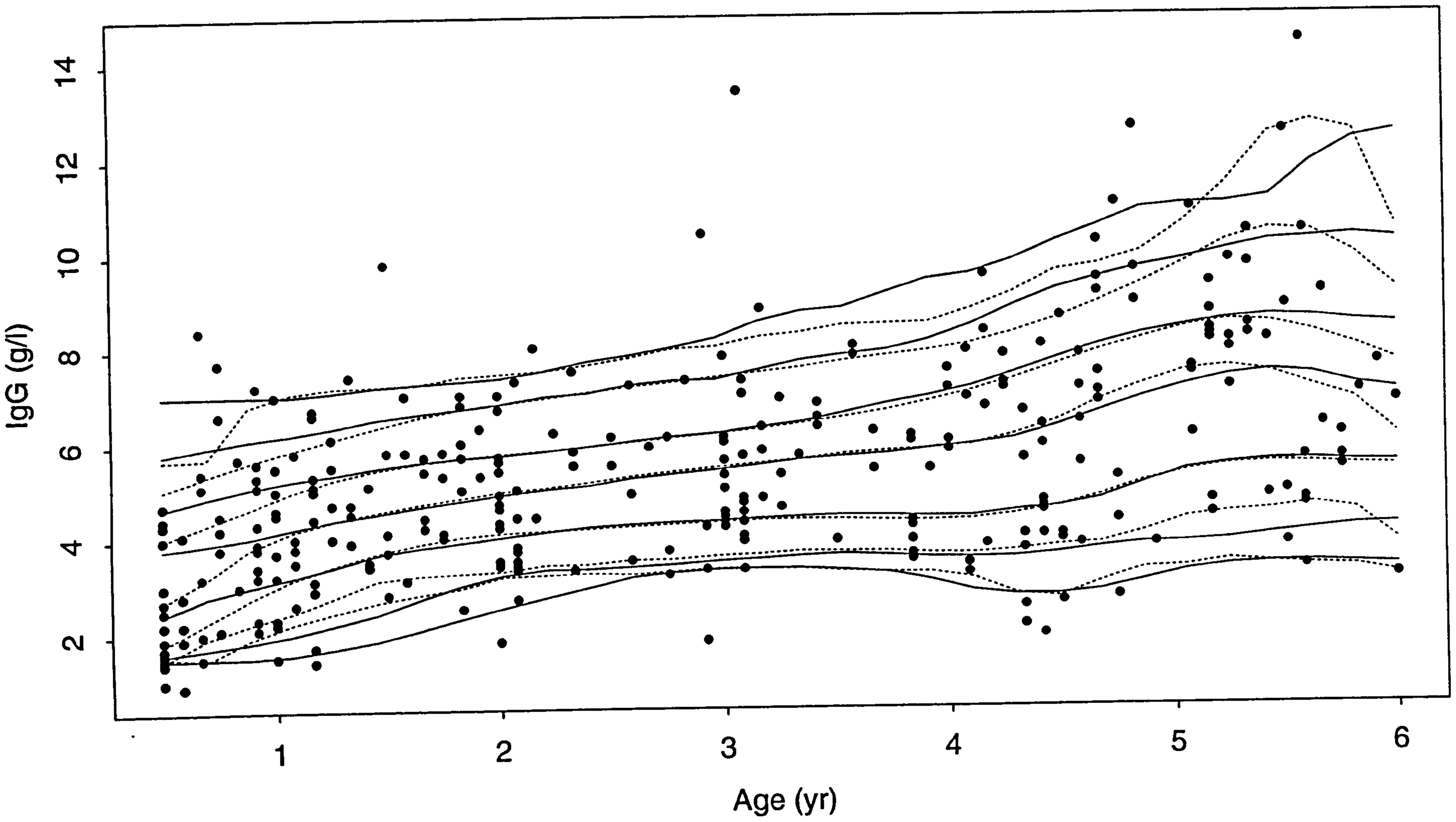


Figure 3.4: Seven smoothed quantile curves for serum data by
local constant fitting (solid lines) and local linear fitting (dotted lines)
5th, 10th, 25th, 50th, 75th 90th and 95th percentiles

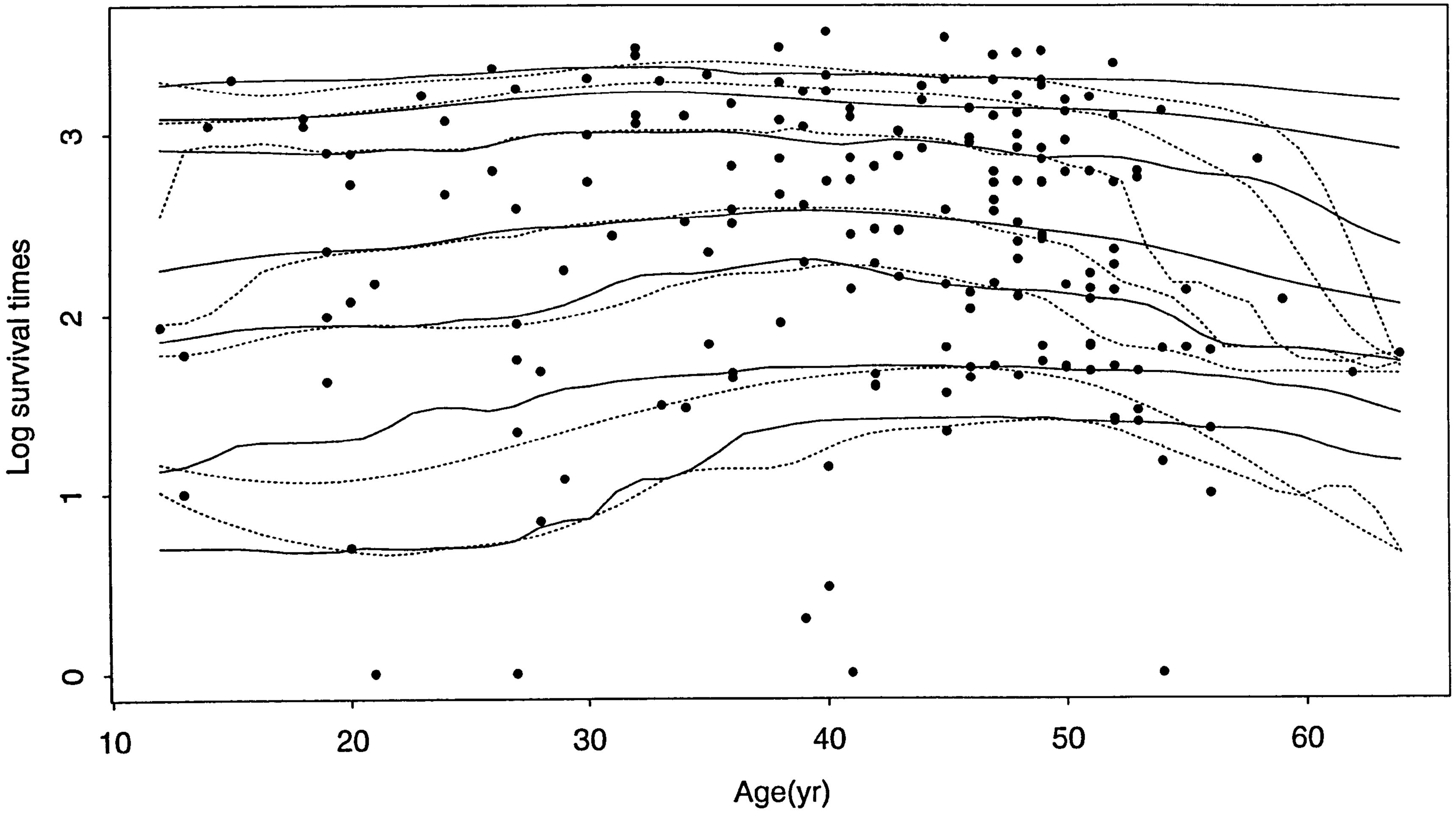


Figure 3.5: Seven smoothed quantile curves for heart transplant data by local constant fitting (solid lines) and local linear fitting (dotted lines) 5th, 10th, 25th, 50th, 75th 90th and 95th percentiles

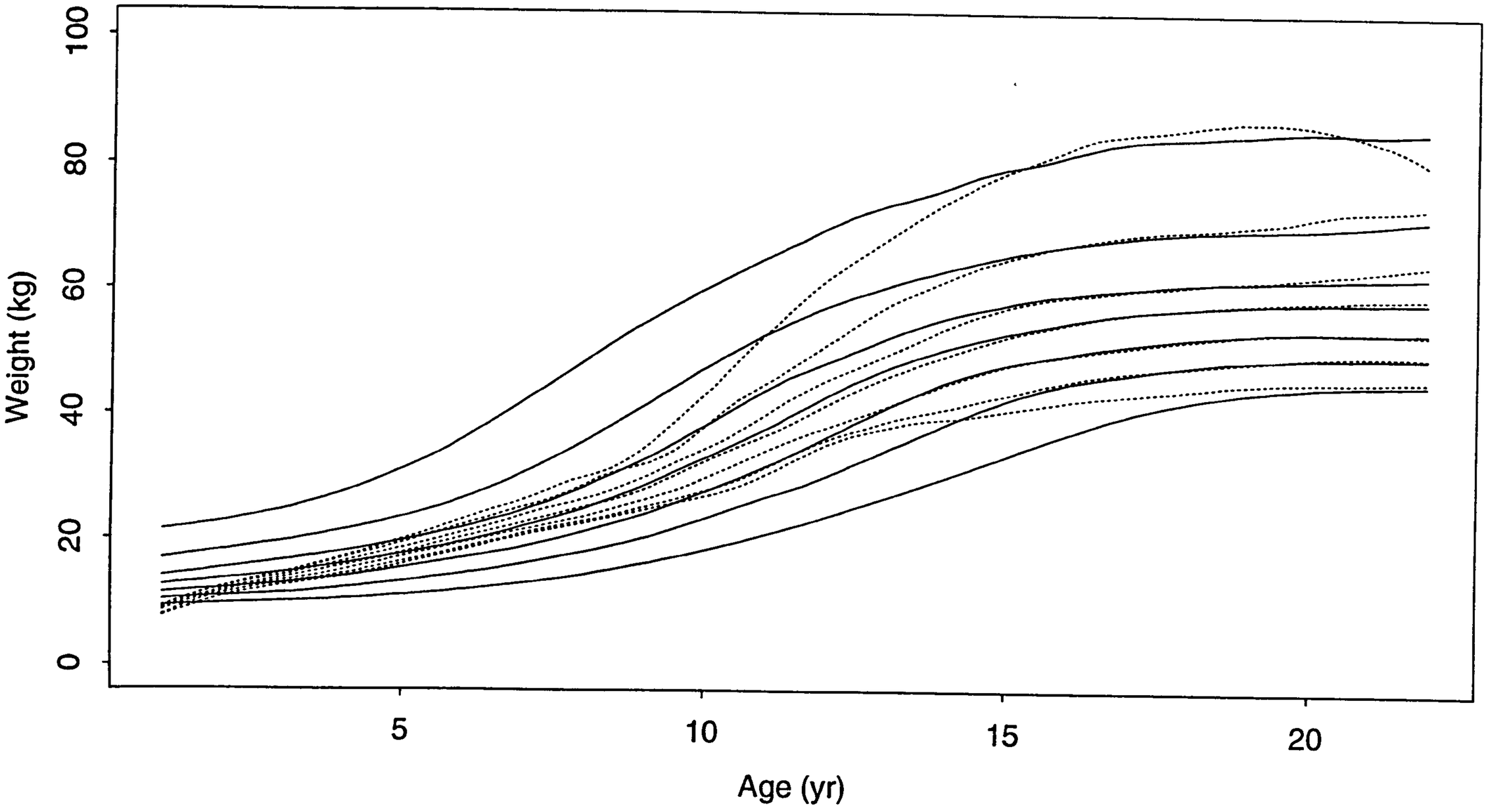


Figure 3.6: Seven smoothed quantile curves for U.S girl data by
local constant fitting (solid lines) and local linear fitting (dotted lines)
3rd, 10th, 25th, 50th, 75th 90th and 97th percentiles

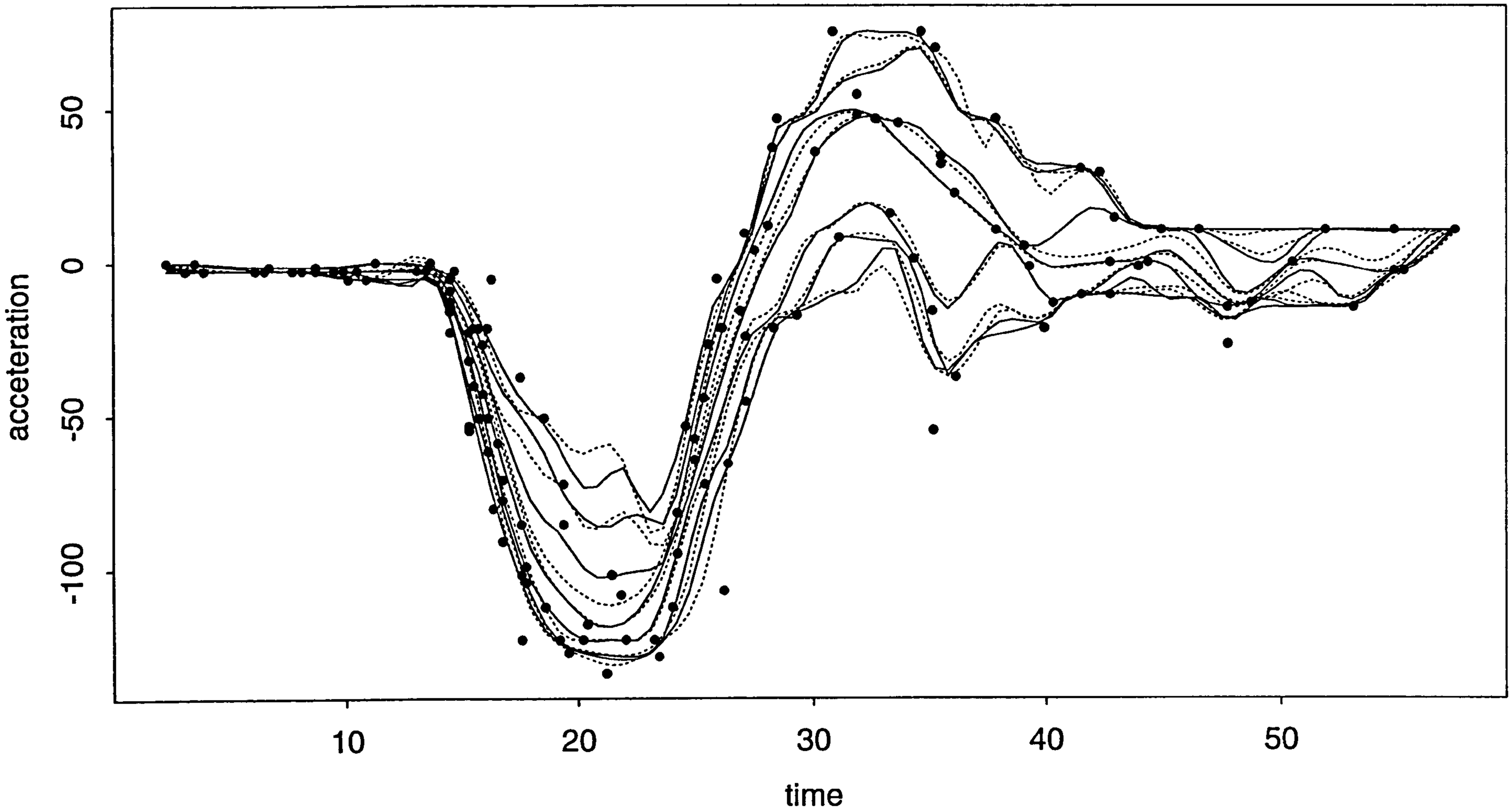
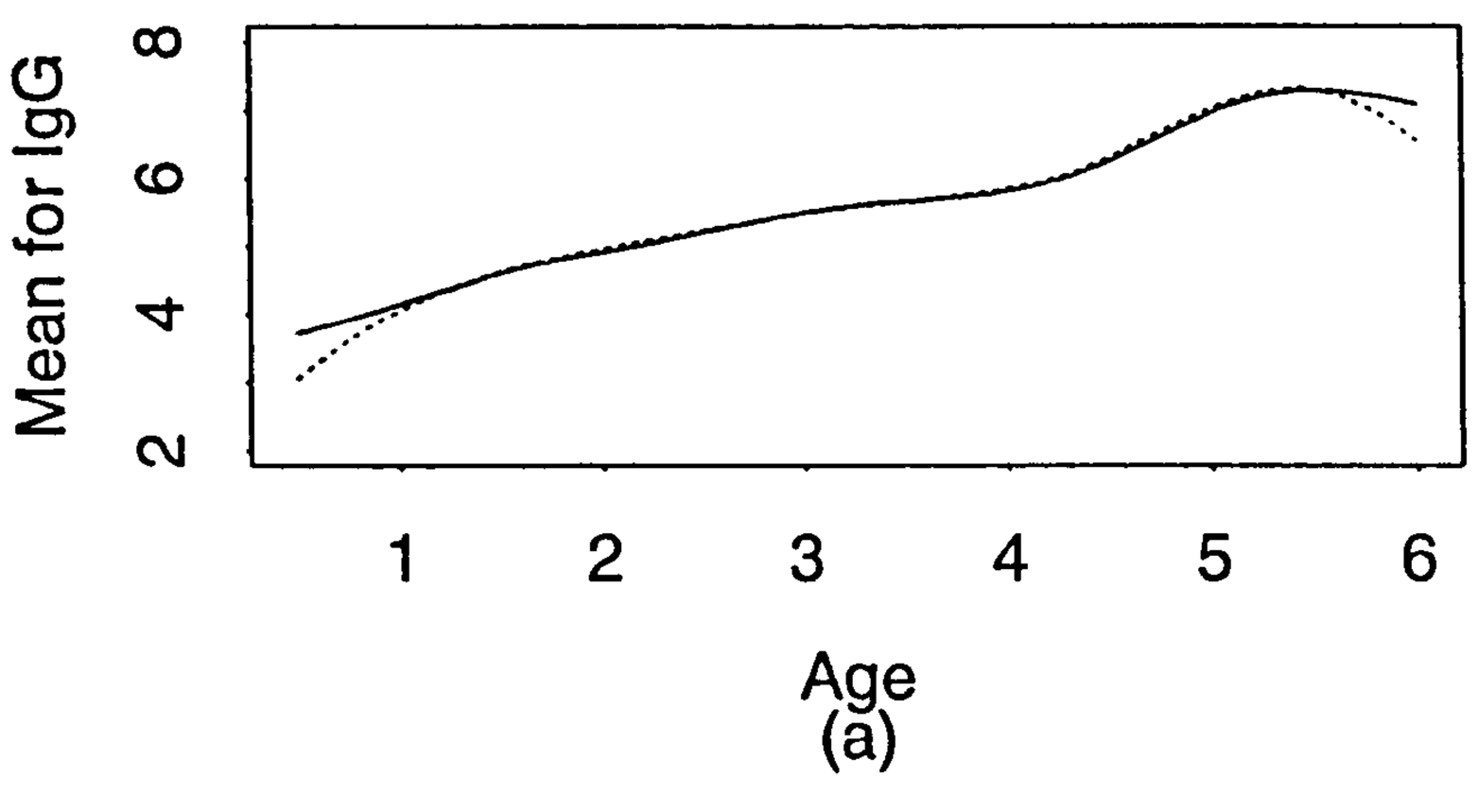


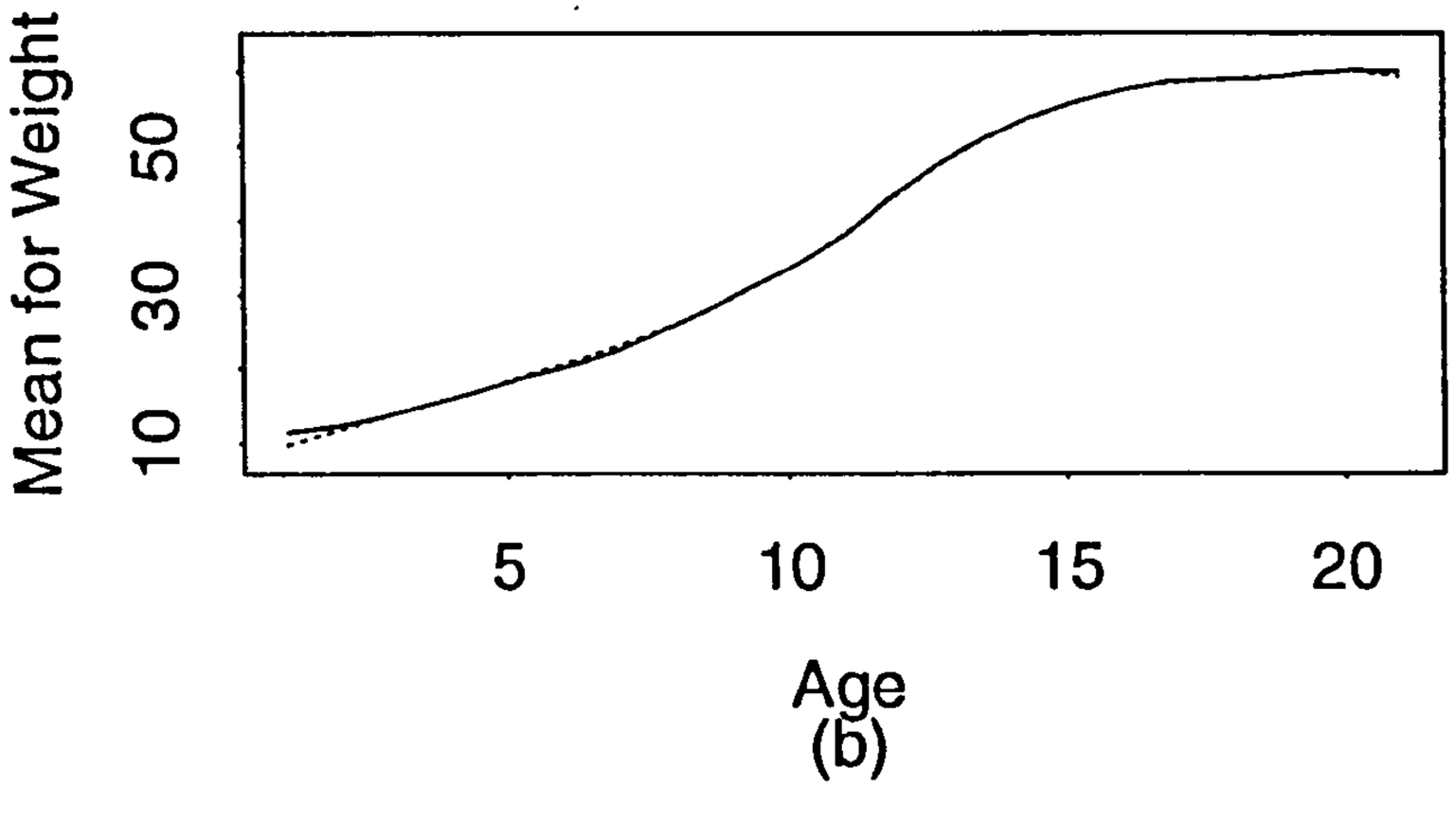
Figure 3.7: Seven smoothed quantile curves for motorcycle data by local constant fitting (solid lines) and local linear fitting (dotted lines) 5th, 10th, 25th, 50th, 75th 90th and 95th percentiles

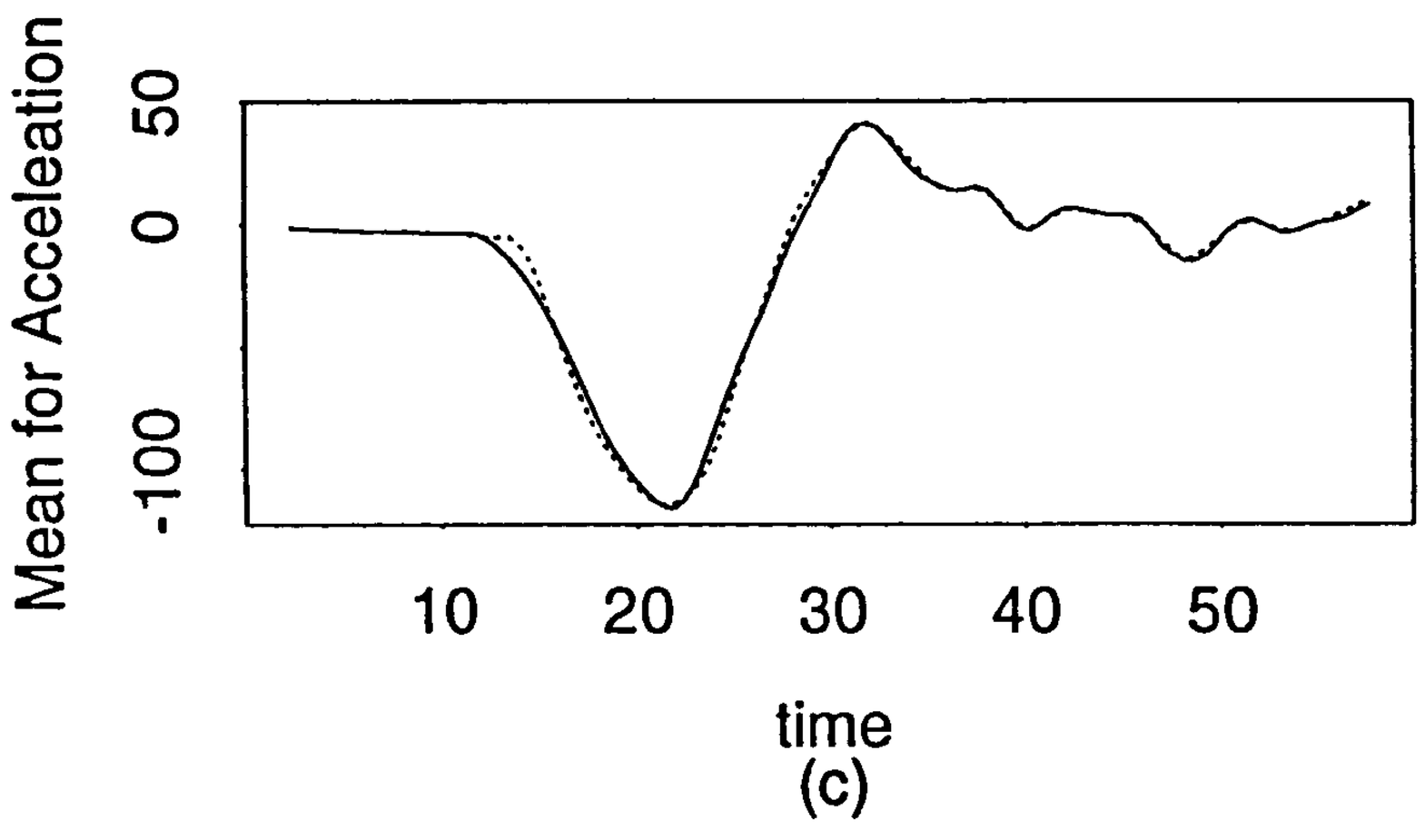
fits in Chapter 2, disappears with local constant fitting, also there is a very considerable effect toward the right in Figure 3.5. It seems appropriate in general to consider the local linear fit as giving the better estimates in these areas although whether the downturns in the smallest quantiles in Figures 3.3 and 3.5 is necessarily appropriate is a moot point.

3.3.2 Estimated Means

The two approaches (local constant and local linear) are also applied on the same data to estimate mean using Ruppert, Sheather & Wand (1995) bandwidth selection technique in local linear case. In the interior, in general, the two estimates are very similar. Only at the boundaries, for serum concentration in Figure 3.8 (a) that appreciable differences emerge. For this reason, almost alone, local linear fitting is superior to local constant. The theory suggests, as in quantile estimation case, that certain combinations of design density and mean slope cause considerable differences. The US girl weight data Figure 3.8 (b), skinfold data Figure 3.8 (d), as well as acceleration data, Figure 3.8 (c) show small differences due to these effects around age 15 and time 15. Also, heart transplant data has same results as Figures 3.8 (b) to (d) although we don't display it here. However, these combinations need to be quite extreme and are fairly rare in practice (see Jones 1995). Therefore, even estimating regression mean, the practical differences between local linear and local constant fitting are not as great as many people suppose.







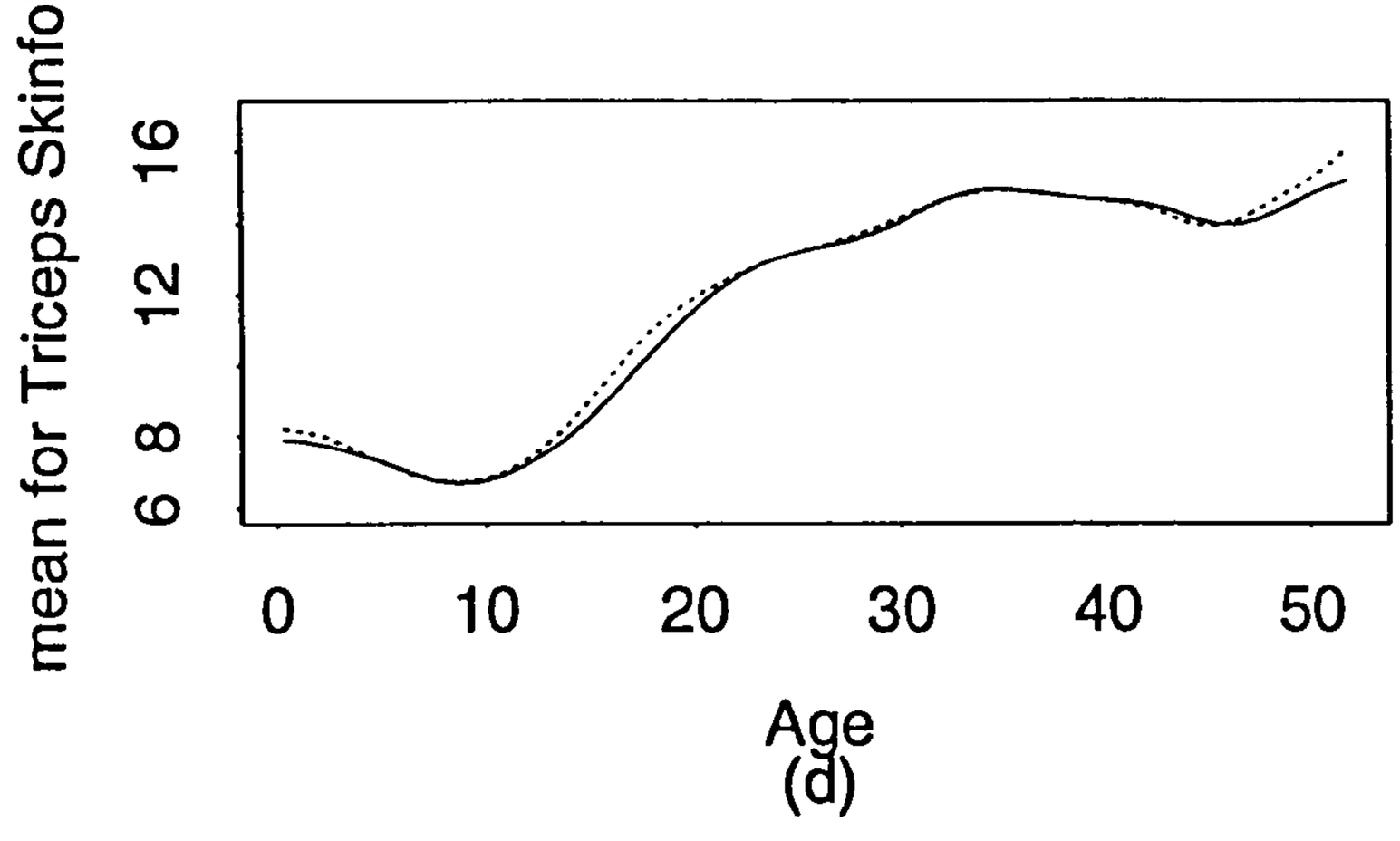


Figure 3.8: Mean fitted by local constant (solid) and local linear (dotted) for data:
(a): Serum, (b): US girl , (c): Acceleration, (d): Gambian women.

3.4 Concluding Remarks

The theoretical analysis and empirical results show that

(1) Local linear fitting has particularly appealing asymptotic mean squared error in terms of intuitive and mathematical simplicity. However, in practice, in comparison with local constant fitting the differences in the interior are not very great.

(2) Boundary influence indeed exists when using local constant fitting in some cases and it is this aspect which clinches the argument in favour of local linear smoothing.

In addition, it is seen that (i) double kernel versions of quantile estimation methodology seem even better (Chapter 2) than the current check function methods, (ii) the bandwidth selection strategy of Chapter 2 seems to work well, (iii) when estimating a quantile, local linear fitting can give quantile derivative estimates as well, although a different bandwidth is needed, and (iv) there is a bigger cost in computing time with local linear than with local constant fitting.

Chapter 4

Relative Efficiency of Double-Kernel Smoothing for Conditional Distributions

4.1 Introduction

Conditional quantiles estimation is closely related to conditional distribution function estimation. The current chapter deals with smoothing problems for conditional distribution function and those aspects related to local linear kernel fitting and conditional quantiles. One hope is that the greater simplicity of the conditional distribution function setting might lead to insight applicable also to conditional quantiles.

Given a pair of random variables (X, Y), the conditional mean $m(x) = E\{Y|X = x\}$ has been the subject of many studies, particularly, its nonparametric smooth-

ing estimation. However, Stone (1977) stressed that the nonparametric estimation of conditional distribution $F(y|x) = P(Y \leq y|X = x)$ is important and merits further investigation.

Since

$$F(y|x) = P(Y \leq y|X = x) = E\{I_{(-\infty,y)}(Y)|X = x\},$$

and a typical estimator of the conditional mean $m(x)$ is of the form $\sum_j w_j(x)Y_j$, then an estimator of $F(y|x)$ is defined as

$$F_{e,n}(y|x) = \sum_j^n w_j(x)I_{(-\infty,y)}(Y) \tag{4.1}$$

where $w_j(x) = w_j(x_1, x_2, .., x_n), 1 \leq j \leq n$, are the weight functions, such that $\sum_j w_j(x) = 1$ and they give higher weight to x_j if x_j is close to x .

The estimator (4.1) defines what is called empirical conditional distribution function, and in a sense it is quite smooth, but further smoothing is a practical advantage.

For $y \in D \subset R$, let $\{X_i, I_{(-\infty,y)}(Y)\}_1^n$ be a new sequence of observations and the choice of weights can be obtained using least-square principle. The response $I_{(-\infty,y)}(Y)$ is the conditional unbiased estimator of $F(y|x)$, however, classical statistics and practical experience shows that an asymptotic unbiased estimator with smaller variance, sometimes, preferable to the unbiased estimators.

Also, the advantage of double-kernel method in smoothing regression quantiles suggests an alternative estimator of $F(y|x)$ of form:

$$F_{d,n}(y|x) = \sum_j^n w_j(x)\Omega(\frac{y - Y_j}{b}) \tag{4.2}$$

where $\Omega(t) = \int_{-\infty}^t W(u)du$, W is a symmetric kernel density function, and b is bandwidth.

Like the estimator (4.1), one can build or re-display many estimating classes for other quantities based on (4.2), such as

1) Conditional density estimating class $\frac{d}{dy}F_{d,n}(y|x) = \sum_j w_j(x)W(\frac{y-Y_j}{b})$ (Fan, Yao and Tong, 1996);

2) Conditional quantiles estimating class $F^{-1}_{d,n}(y|x) = p$ and $0 < p < 1$ (Chapter 2) assuming these are strictly increasing and continuous in x.

3) Conditional moments estimating class: $\hat{m}^r(x) = \int y^r dF_{d,n}(y|x)$ for $r = 1, 2, ...$ assuming that the corresponding rth moment of kernel W exists. In particular when $r = 1$ this defines regression mean estimating class

$$\hat{m}(x) = \int y dF_{d,n}(y|x) = \sum_j w_j(x)Y_j.$$

Further, $b = h$, (4.2) gives a single bandwidth with double-kernel estimator

$$F_{s,n}(y|x) = \sum_j^n w_j(x)\Omega(\frac{y - Y_j}{h}) \tag{4.3}$$

which has an obvious attraction of selecting one bandwidth instead of two in (4.2).

As seen the centre part of these definitions (4.1-4.3) is the weight function w. Without loss of generality, only common local linear weight functions are considered here. Intuitively, a more general estimator, $F_{d,n}(y|x)$ should do as well as either $F_{e,n}(y|x)$ or $F_{s,n}(y|x)$, but it depends on a second bandwidth b. Furthermore, the relationship between $F_{e,n}(y|x)$ and $F_{s,n}(y|x)$ is unknown, although one expects $F_{s,n}(y|x)$ does better than $F_{e,n}(y|x)$. These issues are investigated in the following section.

4.2 Asymptotic Relative Efficiency

To study the properties of the estimators (4.1-4.3), the following notations are adopted to be consistent with (4.1-4.3)

$$\begin{aligned} MSE_e(x,y) &= E\{F_{e,n}(y|x) - F(y|x)\}^2 \\ MSE_d(x,y) &= E\{F_{d,n}(y|x) - F(y|x)\}^2 \\ MSE_s(x,y) &= E\{F_{s,n}(y|x) - F(y|x)\}^2 \end{aligned}$$

and

$$\begin{aligned} \mu_2(K) &= \int t^2 K(t)dt \\ \mu_2(W) &= \int t^2 W(t)dt \\ R(K) &= \int K^2(t)dt \\ \alpha(W) &= \int \Omega(t)(1-\Omega(t))dt \end{aligned}$$

then

Theorem 4.1: If $h \to 0$, $b \to 0$ and local linear kernel weight fit, the asymptotic mean square error of the estimators 4.1 -4.3 are

$$\begin{aligned} AMSE_e(x,y) &= 1/4\Big(F^{2,0}(y|x)\mu_2(K)\Big)^2 h^4 \\ &+ \frac{R(K)}{ng(x)h}F(y|x)(1-F(y|x)) \qquad (4.4) \\ AMSE_d(x,y) &= 1/4\Big(F^{2,0}(y|x)\mu_2(K)h^2 + F^{0,2}(y|x)\mu_2(W)b^2\Big)^2 \\ &+ \frac{R(K)}{ng(x)h}F(y|x)(1-F(y|x)) \\ &- \frac{1}{nhg(x)}R(K)\alpha(W)f(y|x)b \qquad (4.5) \\ AMSE_s(x,y) &= 1/4\Big(F^{2,0}(y|x)\mu_2(K) + F^{0,2}(y|x)\mu_2(W)\Big)^2 h^4 \\ &+ \frac{R(K)}{ng(x)h}F(y|x)(1-F(y|x) \end{aligned}$$

$$- \quad \frac{1}{n} R(K)\alpha(W)\frac{f(y|x)}{g(x)} \tag{4.6}$$

Here AMSE stands for asymptotic mean square error.

The proof of Theorem 4.1 is easy. In fact, $AMSE_e(x,y)$ can be derived from the mean square error of local linear kernel fit regression mean $m(x)$ with $m(x) = E\{I_{(-\infty,y)}(Y)|X=x\} = F(y|x)$ and $\sigma^2(x) = Var\{I_{(-\infty,y)}(Y)|X=x\} = F(y|x)(1-F(y|x))$. Analogously, $AMSE_d(x,y)$ can be proved along the line of proving the Theorem 2.1 of Chapter 2, while $AMSE_s(x,y)$ is a special case of $AMSE_d(x,y)$.

From Theorem 4.1, the asymptotic mean square error of estimator in (4.1) and (4.2) are equal as $b \to 0$, i.e $AMSE_d(x,y) \to AMSE_e(x,y)$. Also, for any $b \geq 0$, if $F^{0,2}(y|x) = \frac{d}{dy}f(y|x) = f'(y|x) \approx 0$, the asymptotic biases of the three estimators are dominated by $F^{2,0}(y|x)$, and both $F_{s,n}(y|x)$ and $F_{d,n}(y|x)$ have possible smaller variance than $F_{e,n}(y|x)$.

To investigate further the relative efficiency, in terms of MISE, let us firstly compare $AMSE_e(x,y)$ and $AMSE_s(x,y)$ in terms of the global measurements $\int\int\{AMSE_e(x,y)\}dxdy$ and $\int\int\{AMSE_s(x,y)\}dxdy$. Then Theorem 4.2 is true:

Theorem 4.2: Suppose $F^{2,0}(y|x)$ and $F^{0,2}(y|x)$ are continuous about x and y, then the global properties defined by $\int\int\{AMSE(x,y)\}dxdy$ of $F_{e,n}(y|x)$ and $F_{s,n}(y|x)$ are:

i) In terms of bias, single kernel with single bandwidth estimator $F_{e,n}(y|x)$ is better than double kernel with single bandwidth estimator $F_{s,n}(y|x)$.

ii) In terms of variance, double kernel with single bandwidth estimator $F_{s,n}(y|x)$ is better than single kernel with single bandwidth estimator $F_{e,n}(y|x)$.

Proof: ii) is obvious from Theorem 4.1 and $\int\int \frac{1}{n}R(K)\alpha(W)\int \frac{f(y|x)}{g(x)}dxdy > 0$ for all n and kernel K and W.

i) is true as $\int\int F^{2,0}(y|x)F^{0,2}(y|x)dxdy \geq 0$. In fact, integration by parts twice (firstly integration by parts about y then x),

$$\int\int F^{2,0}(y|x)F^{0,2}(y|x)dxdy = \int\int \{F^{1,1}(y|x)\}^2 dxdy \geq 0.$$

Remark: Asymptotically bias compares first order, so single kernel with single bandwidth seems to do better, but variance reduction in practice is also important, so $F_{s,n}(y|x)$ and $F_{d,n}(y|x)$ remain difficult to compare, and this can still be seen from their best possible asymptotic MISEs about x with weight function $v(x)$:

$$\begin{aligned} AMISE_e^* &= \frac{5}{4}n^{-4/5}(\mu_2(K))^{2/5}R(K)^{4/5}(\int\{F^{2,0}(y|x)\}^2 v(x)dx)^{1/5} \\ &\quad (\int \frac{F(y|x)(1-F(y|x)}{g(x)}v(x)dx)^{4/5} \\ AMISE_s^* &= \frac{5}{4}n^{-4/5}R(K)^{4/5}\left(\int(\mu_2(K)F^{2,0}(y|x)+\mu_2(W)F^{0,2}(y|x))^2 v(x)dx\right)^{1/5} \\ &\quad (\int \frac{F(y|x)(1-F(y|x))}{g(x)}v(x)dx)^{4/5} \\ &\quad - \frac{1}{n}R(K)\alpha(W)\int\frac{f(y|x)}{g(x)}v(x)dx \end{aligned}$$

Similarly, Theorem 4.3 compares relative efficiency of the double kernel with double bandwidth estimator $F_{d,n}(y|x)$ with respect to $F_{e,n}(y|x)$ and $F_{s,n}(y|x)$ in terms of $\int\int AMSE(x,y)dxdy$.

Theorem 4.3: Under the same conditions of Theorem 4.2 and if positive $\frac{n}{R(K)\alpha(W)}b$ is selected according to

$$min\{\frac{1}{h^3}\frac{\int\int\frac{f(y|x)}{g(x)}dxdy}{\mu_2(K)\mu_2(W)\int\int F^{2,0}(y|x)f'(y|x)dxdy}, \left(\frac{2}{\mu_2(W)^2 h}\frac{\int\int\frac{f(y|x)}{g(x)}dxdy}{\int\int(f'(y|x))^2 dxdy}\right)^{1/3}\}$$

while h minimizes $\int\int AMISE_e(x,y)dxdy$, then

$$\int\int AMSE_d(x,y)dxdy \leq \int\int AMSE_e(x,y)dxdy,$$

$$\int\int AMSE_d(x,y)dxdy \leq \int\int AMSE_s(x,y)dxdy.$$

Remark: Usually, we can select b according to

$$b = \frac{R(K)\alpha(W)}{n}\left(\frac{2}{\mu_2(W)^2 h}\frac{\int\int \frac{f(y|x)}{g(x)}dxdy}{\int\int (f'(y|x))^2 dxdy}\right)^{1/3},$$

because of $h \sim n^{-1/5}$ when n is big enough.

Proof: When b is selected according to the minimizer of both, then

$$\frac{1}{4}\int\int (f'(y|x))^2 dxdy\mu_2(W)^2 b^4 + \frac{\mu_2(K)\mu_2(W)}{2}\int\int F^{2,0}(y|x)f'(y|x)dxh^2b^2$$

$$-b\frac{R(K)\alpha(W)}{nh}\int\frac{f(y|x)}{g(x)}v(x)dx \leq 0,$$

so result follows.

Note that this is a "second order effect" relying on negative second variance term.

These asymptotic results indeed show that $F_{d,n}(y|x)$ is the most efficient for estimating $F(y|x)$ among three estimators (as must be the case since $F_{e,n}(y|x)$ and $F_{s,n}(y|x)$ are special cases of $F_{d,n}(y|x)$). But no absolute conclusion is drawn between $F_{e,n}(y|x)$ and $F_{s,n}(y|x)$.

4.3 Exact Relative Efficiency

As seen it is difficult to obtain exact and explicit expressions for the integrated mean square errors MISE, even in mixtures of normal distributions. To demon-

strate this, consider a regression model

$$Y = m(X) + \epsilon,$$

with covariate variable X independent of random error ϵ, and $X \sim g(x)$, $\epsilon \sim \chi(z)$, then the conditional density of Y given $X = x$ is $f(y|x) = \chi(y - m(x))$. The value of $f(y|x)$ at pth quantile $q_p(x)$ can be obtained from $f(q_p(x)|x) = \chi(z_p)$. Thus the estimate of conditional distribution is based on the model and mainly determinated by the error distribution $\chi(z)$ and not the regression component $m(x)$.

Consider the integrated MSE about x, of $F_{e,n}(y|x)$, for example, defined as

$$MISE(y) = E \int \{F_{e,n}(y|x) - F(y|x)\}^2 dx$$

Let $MISE_e(y)$, $MISE_d(y)$ and $MISE_s(y)$ be, respectively, the MISE of single-kernel, double-kernel with double-bandwidth, and double-kernel with single-bandwidth and suppose that

$$\begin{aligned} m(x) &= xexp(-x^2) \\ g(x) &\sim N(0,1) \end{aligned}$$

Moreover, a list of distributions extracted from Marron and Wand (1992) is given in Table 4.1 and the graphs of these are displayed in Figures 4.1-4.8; these will serve the role of χ in our simulations. The integrated square errors (ISE) of the estimator are computed using

$$ISE(y) = \int_{-3}^{3} (estimator - F(y|x))^2 dx,$$

with a equal weight and taking p-quantile of Y as y, and all bandwidths are chosen, with knowledge of the model, to minimize AMISE. Sometimes, in double-kernel with double-bandwidth estimator approximate solution for (h, b) are obtained as either real-valued solution does not exist or it is not unique.

Density	Normal mixture densities
No.1 Gaussian	$N(0,1)$
No.2 Skewed unimodal	$\frac{1}{5}N(0,1) + \frac{1}{5}N(\frac{1}{2},(\frac{2}{3})^2) + \frac{3}{5}N(\frac{13}{12},(\frac{2}{3})^2)$
No.3 Kurtotic unimodal	$\frac{2}{3}N(0,1) + \frac{1}{3}N(0,(\frac{1}{10})^2)$
No.4 Outlier	$\frac{1}{10}N(0,1) + \frac{9}{10}N(0,(\frac{1}{10})^2$
No.5 Bimodal	$\frac{1}{2}N(-1,(\frac{2}{3})^2) + \frac{1}{2}N(1,(\frac{2}{3})^2)$
No.6 Separated bimodal	$\frac{1}{2}N(-\frac{3}{2},(\frac{1}{2})^2) + \frac{1}{2}N(\frac{3}{2},(\frac{1}{2})^2)$
No.7 Skewed bimodal	$\frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2},(\frac{1}{3})^2)$
No.8 Trimodal	$\frac{9}{20}N(-\frac{6}{5},(\frac{3}{5})^2) + \frac{9}{20}N(\frac{6}{5},(\frac{3}{5})^2) + \frac{1}{10}N(0,(\frac{1}{4})^2)$

Table 4.1: Eight normal mixture distributions of ϵ

For comparison of the three estimators 4.1-4.3 define the relative efficiency of an estimator

$$RE = \left(\frac{MISE_d}{MISE_e \ \ or \ \ MISE_s}\right)^{(5/4)}.$$

Data are simulated (simulating x each time) from the distributions in Table 4.1 and for $n = 1000$ and for each estimator the $MISE$'s are calculated for y's corresponding to $p = 0.5$, $p = 0.9$ and $p = 0.1$ and using local linear kernel fitting and standard normal as kernel functions. These results based on 100 simulations are presented in Tables 4.2-4.9 with 3 decimal places.

p	RE (%) of $F_{e,n}(q_p\|x)$	RE (%) of $F_{s,n}(q_p\|x)$
$p = 0.5$	0.240	0.697
$p = 0.9$	0.827	0.980
$p = 0.1$	0.896	0.980

Table 4.2: RE with Gaussian distribution

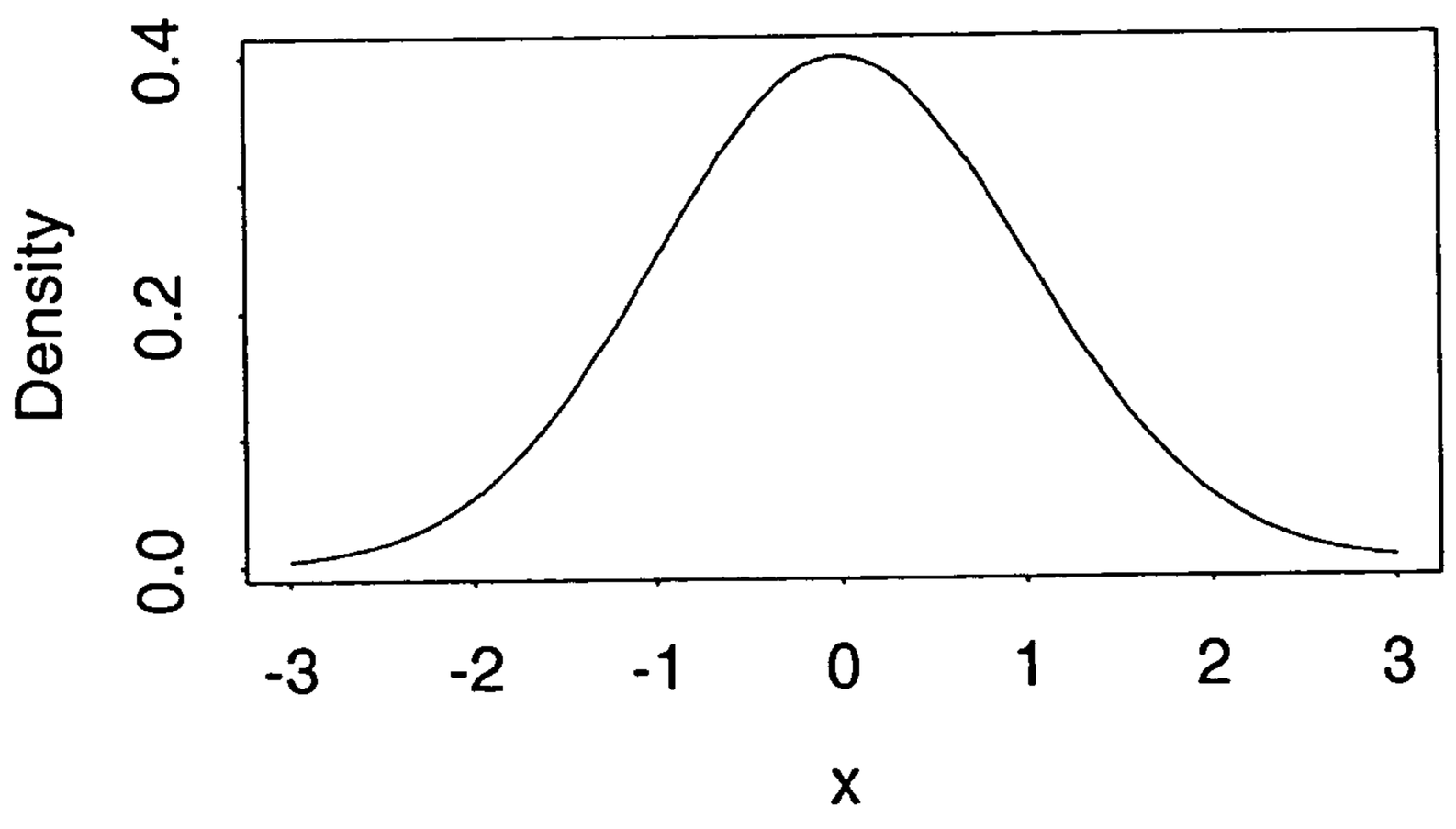


Figure 4.1: Gaussian

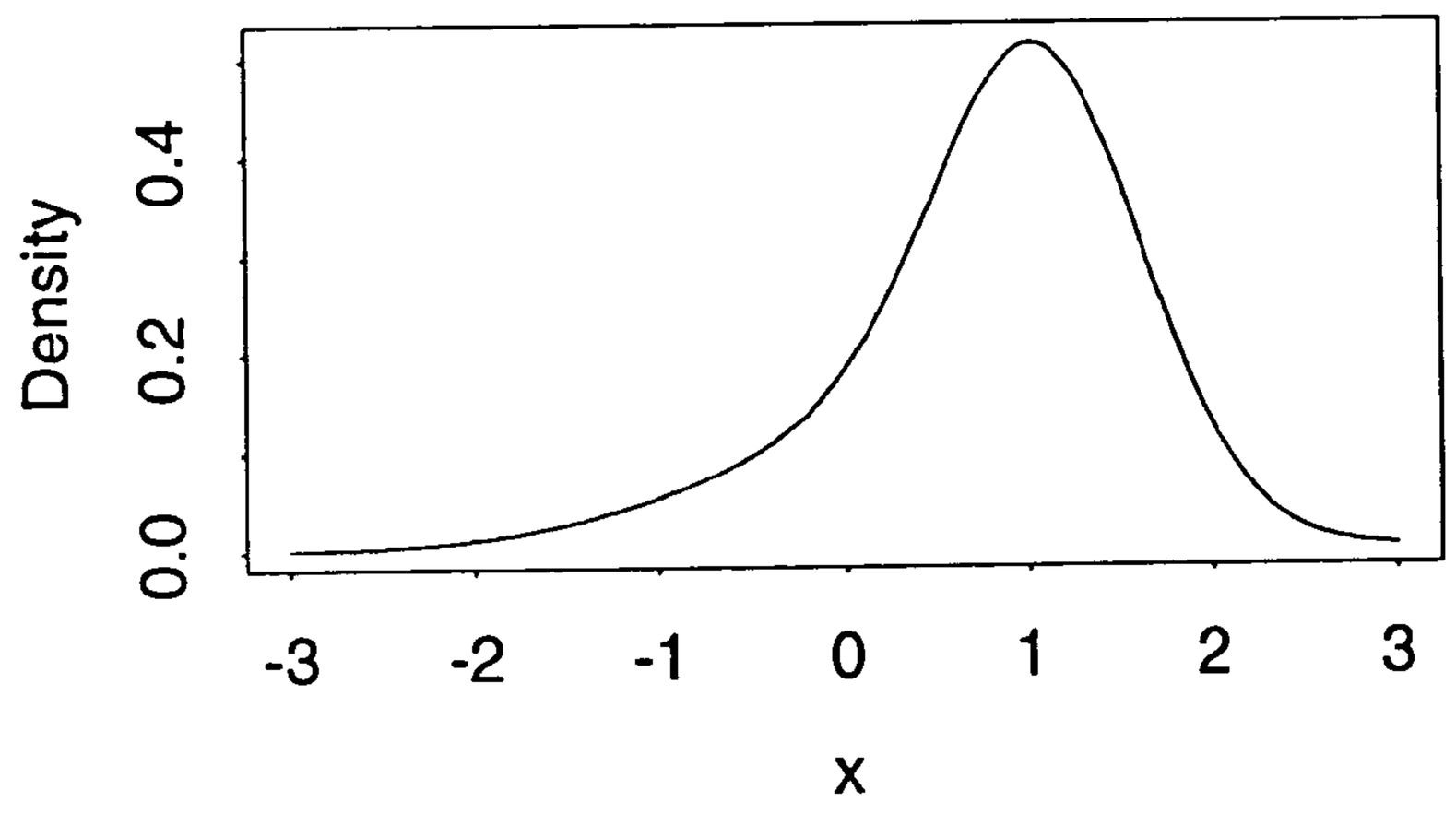


Figure 4.2: Skewed unimodal

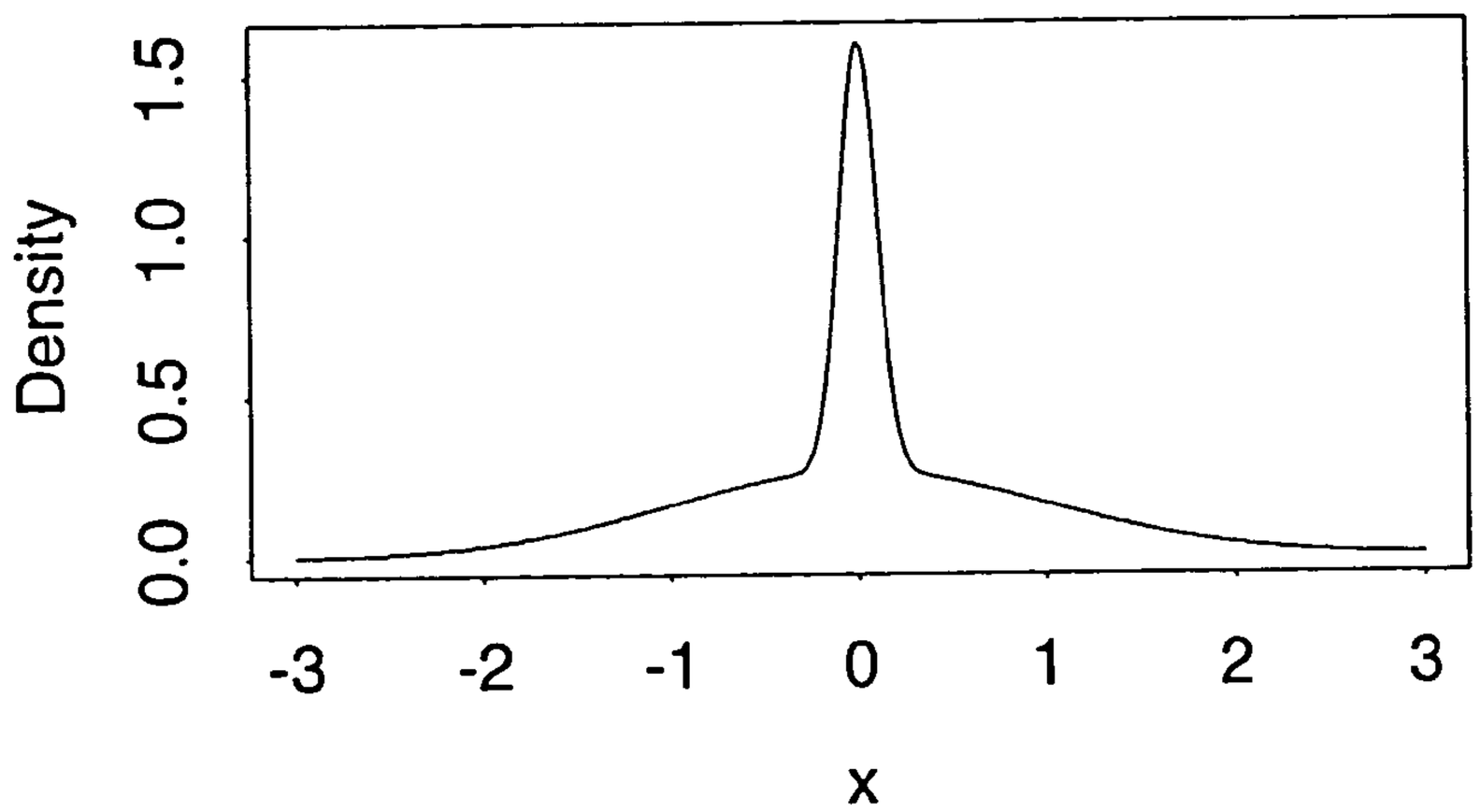


Figure 4.3:Kurtotic unimodal

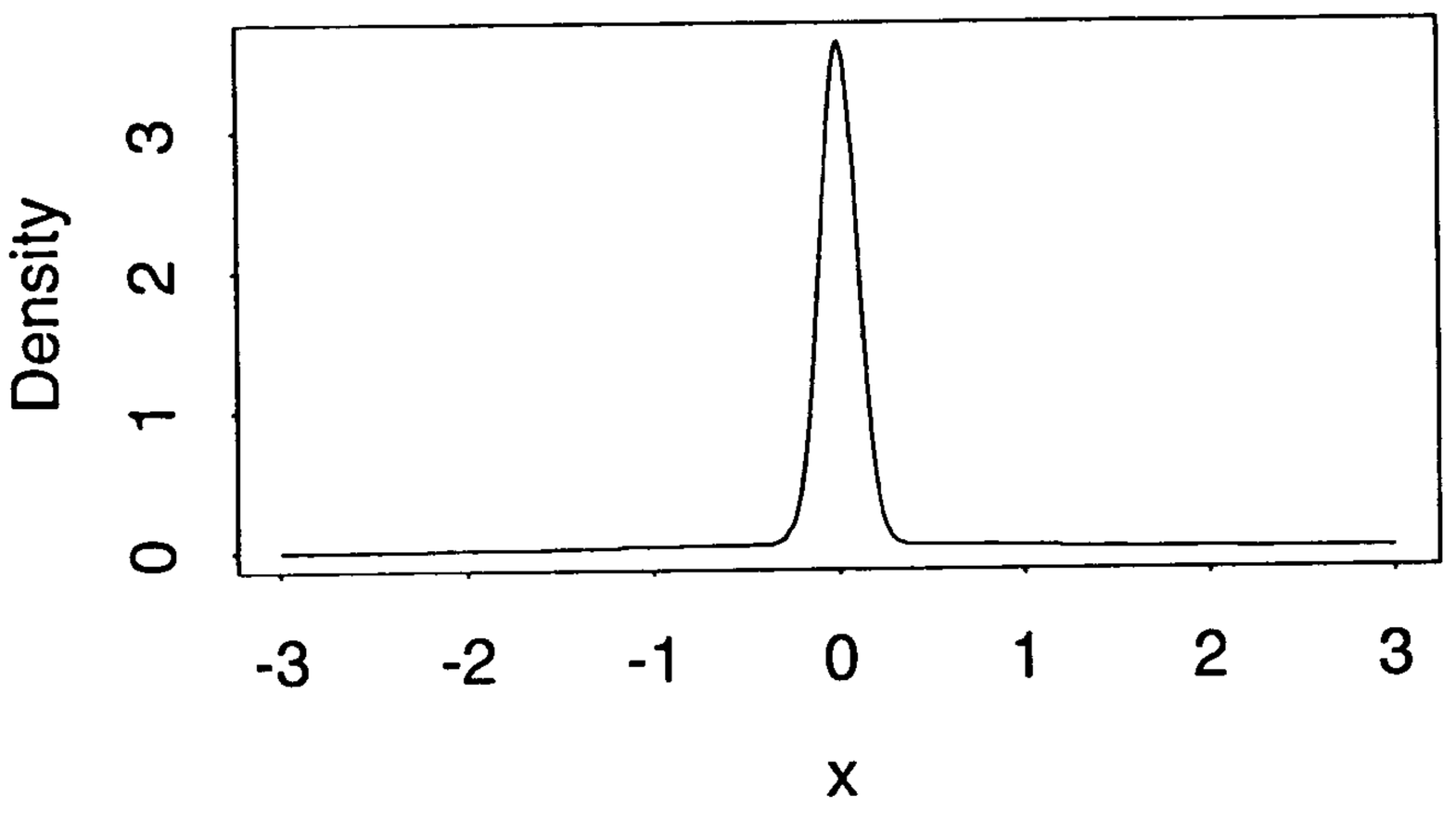
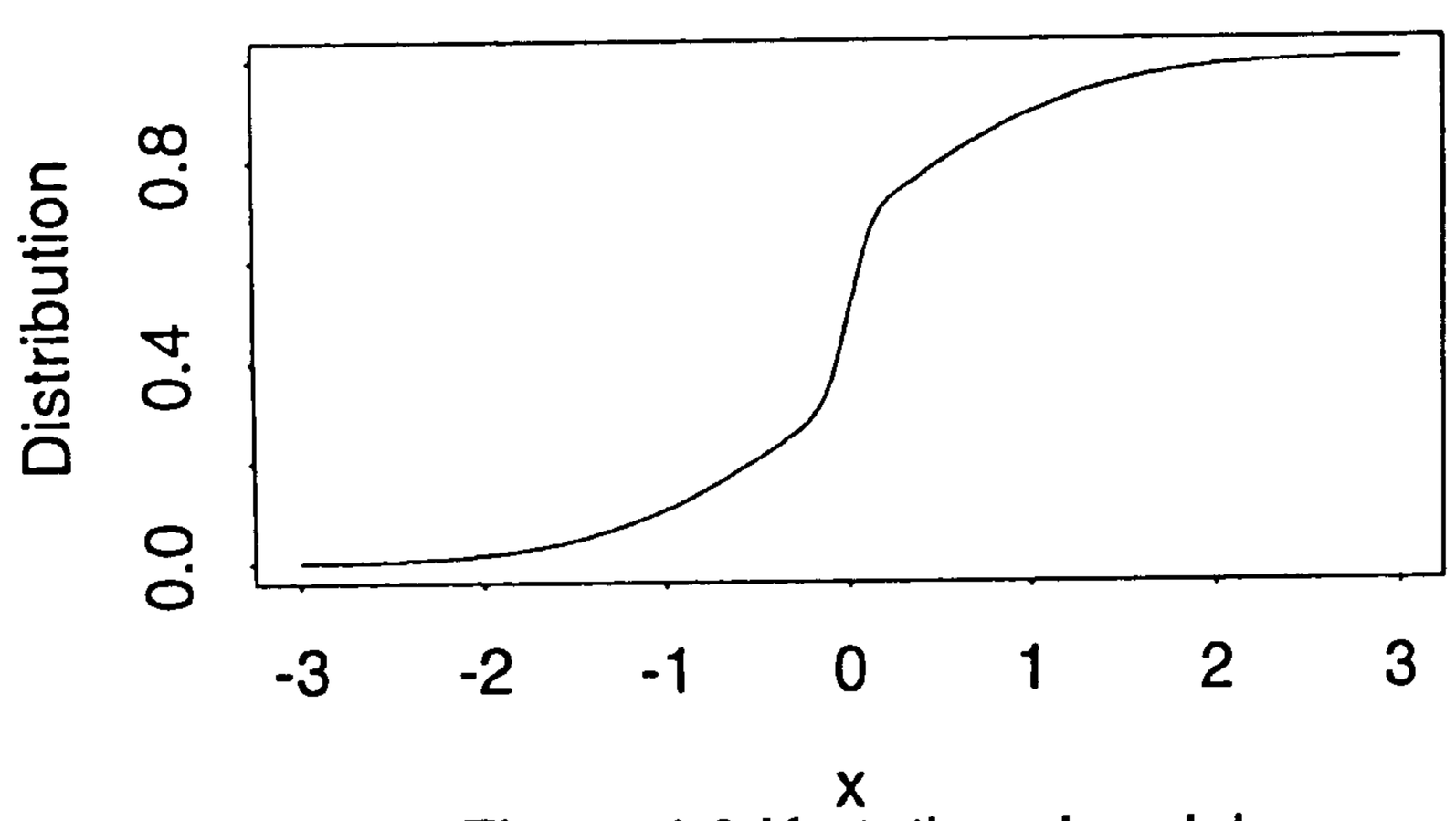
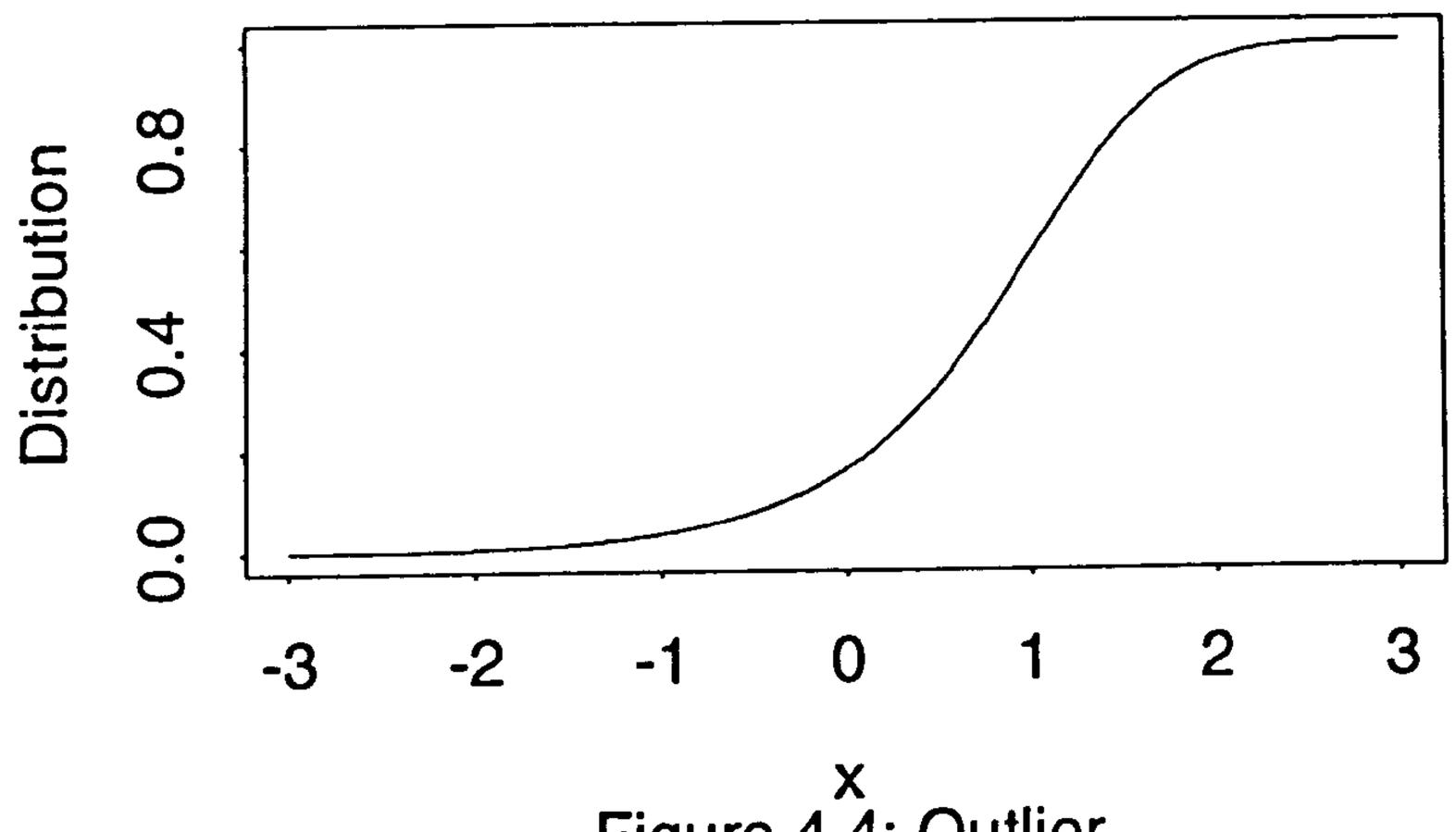


Figure 4.4: Outlier





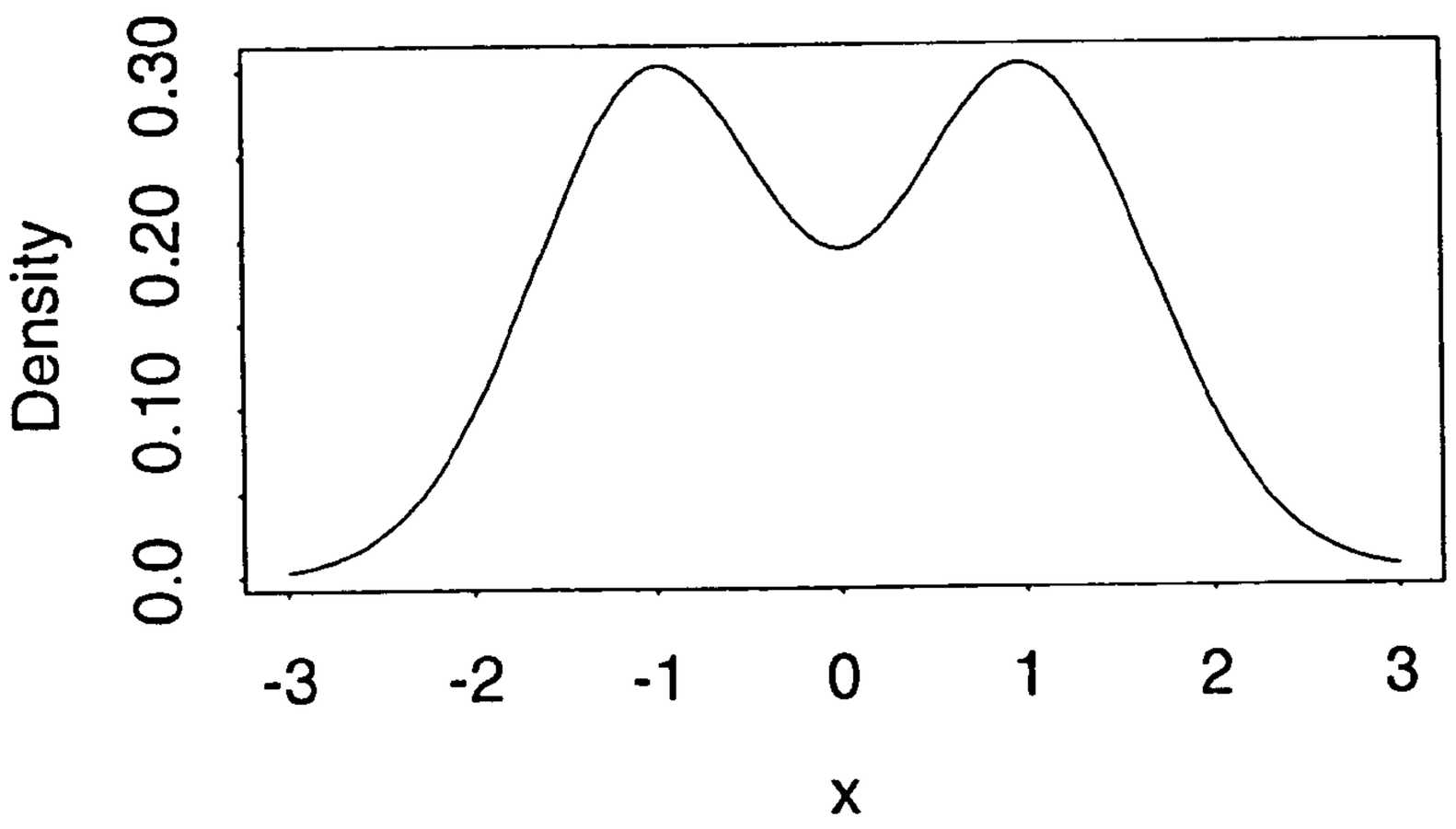
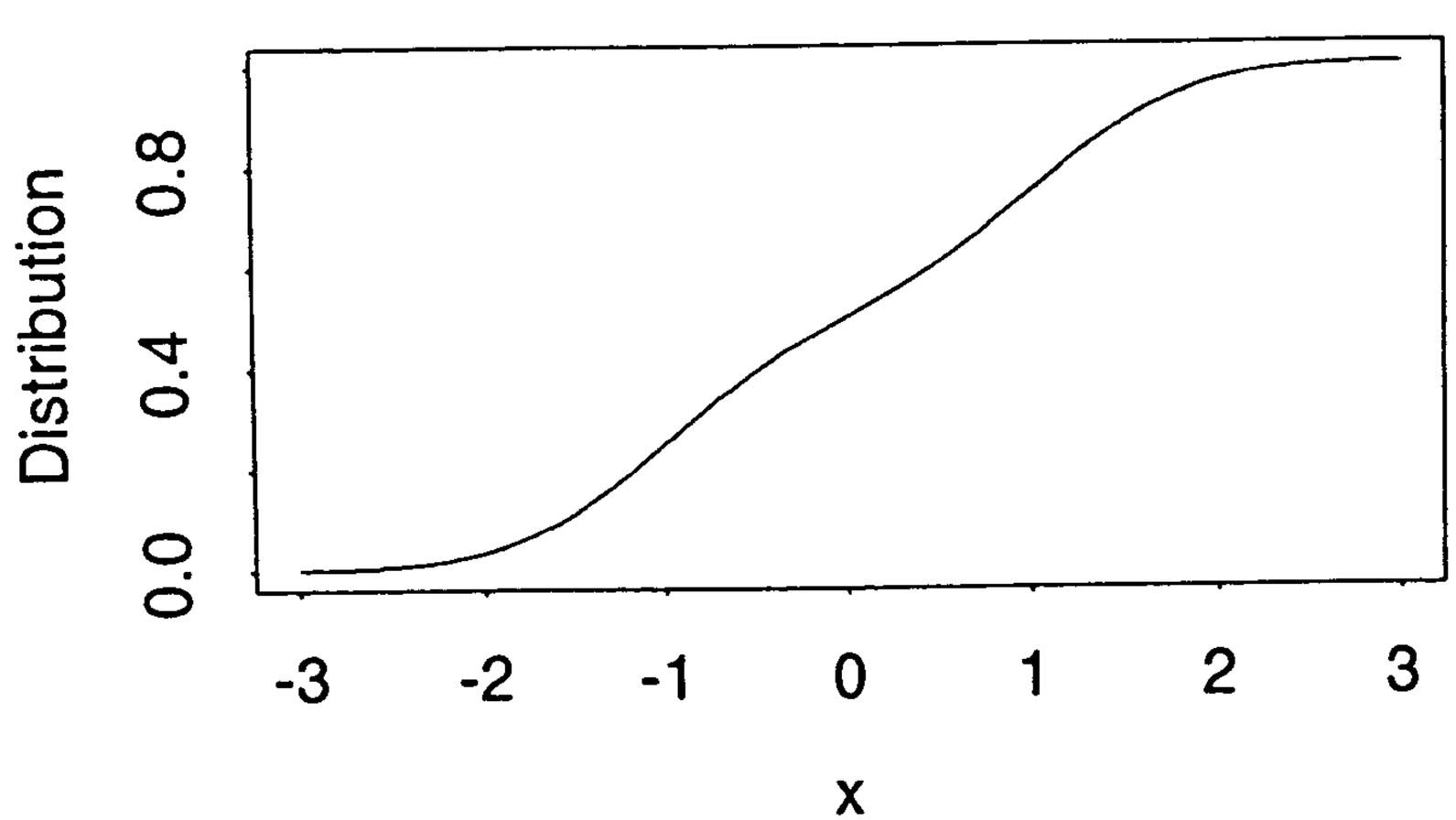


Figure 4.5: Bimodal



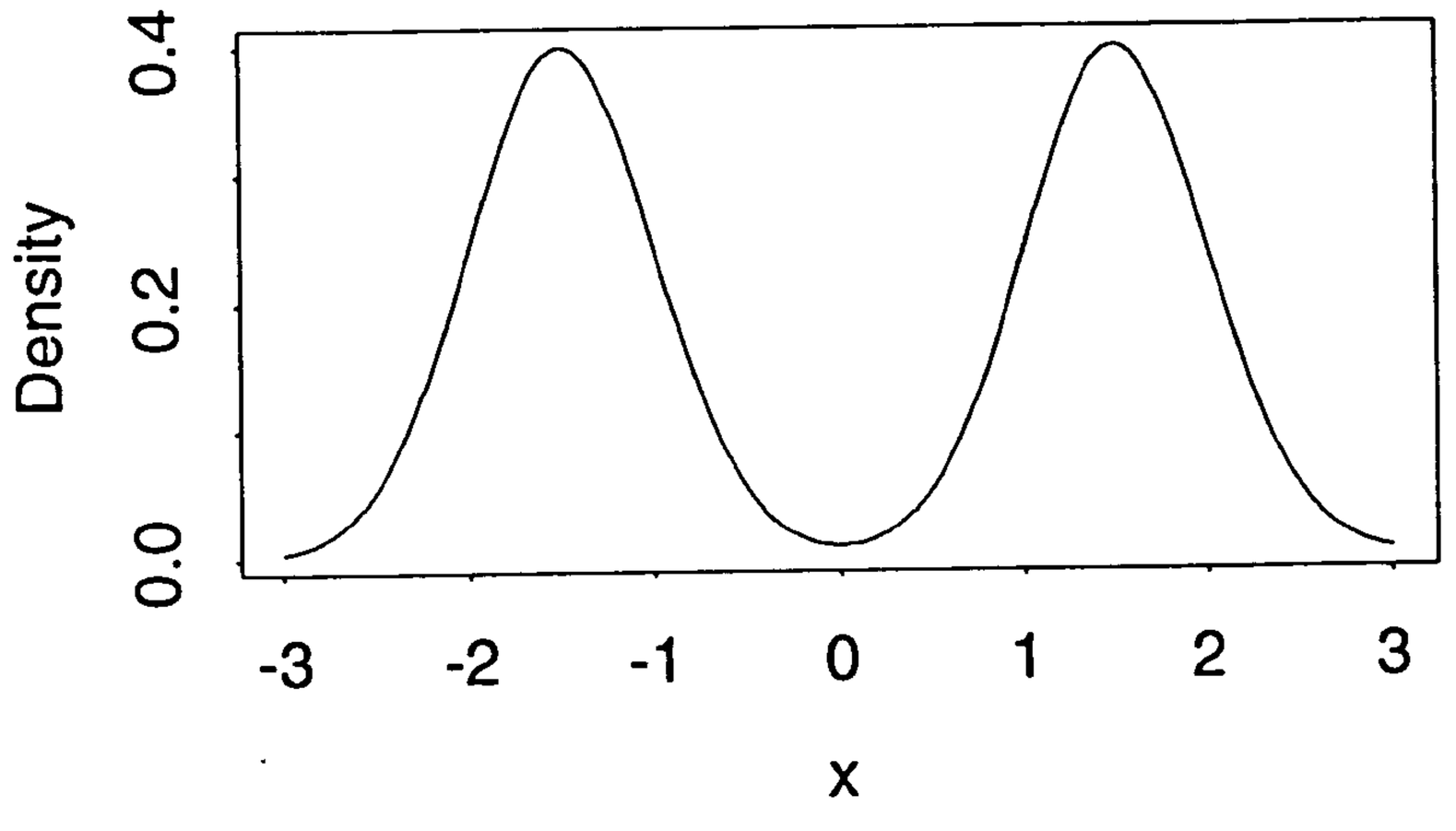


Figure 4.6: Separated bimodal

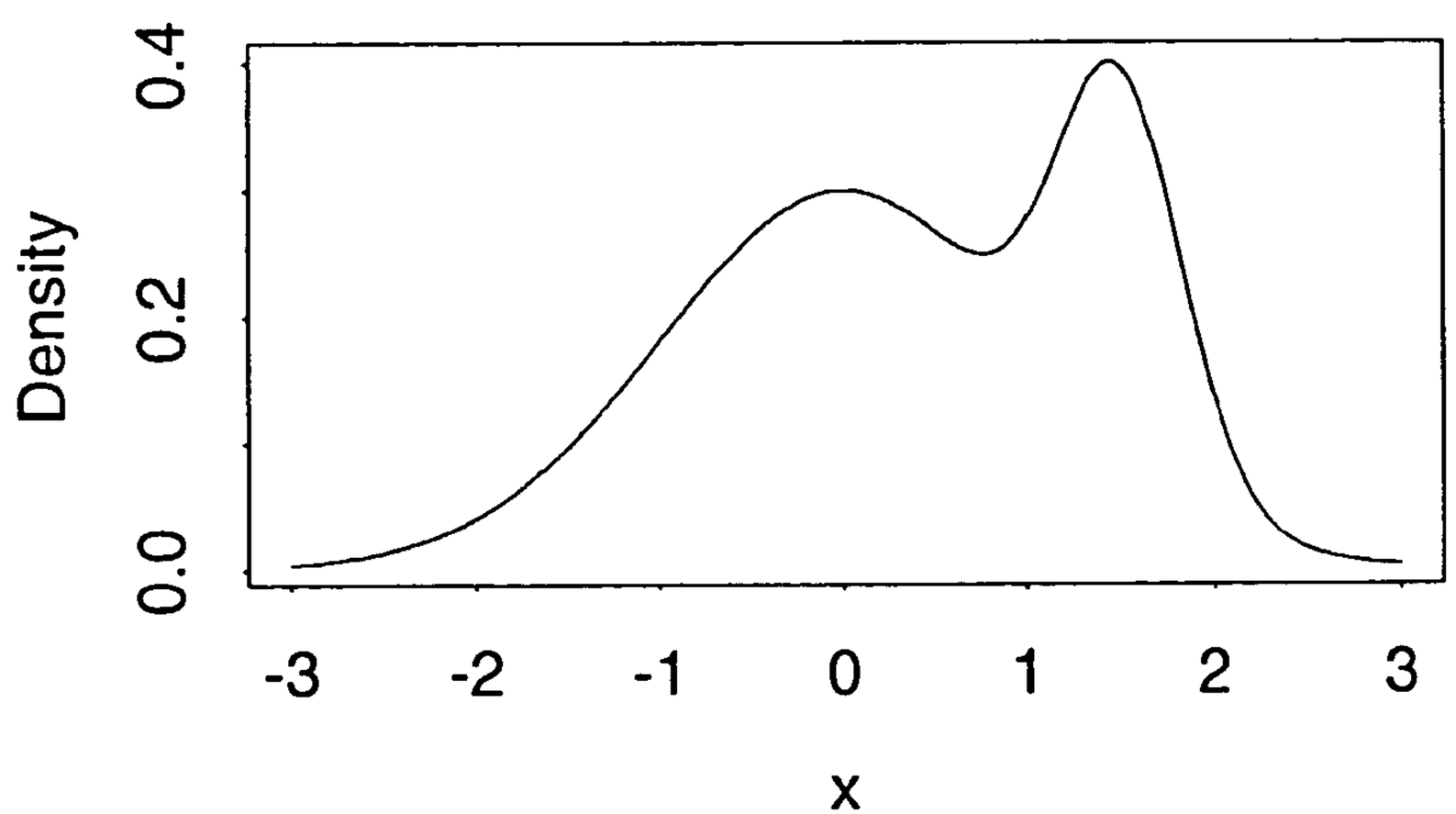
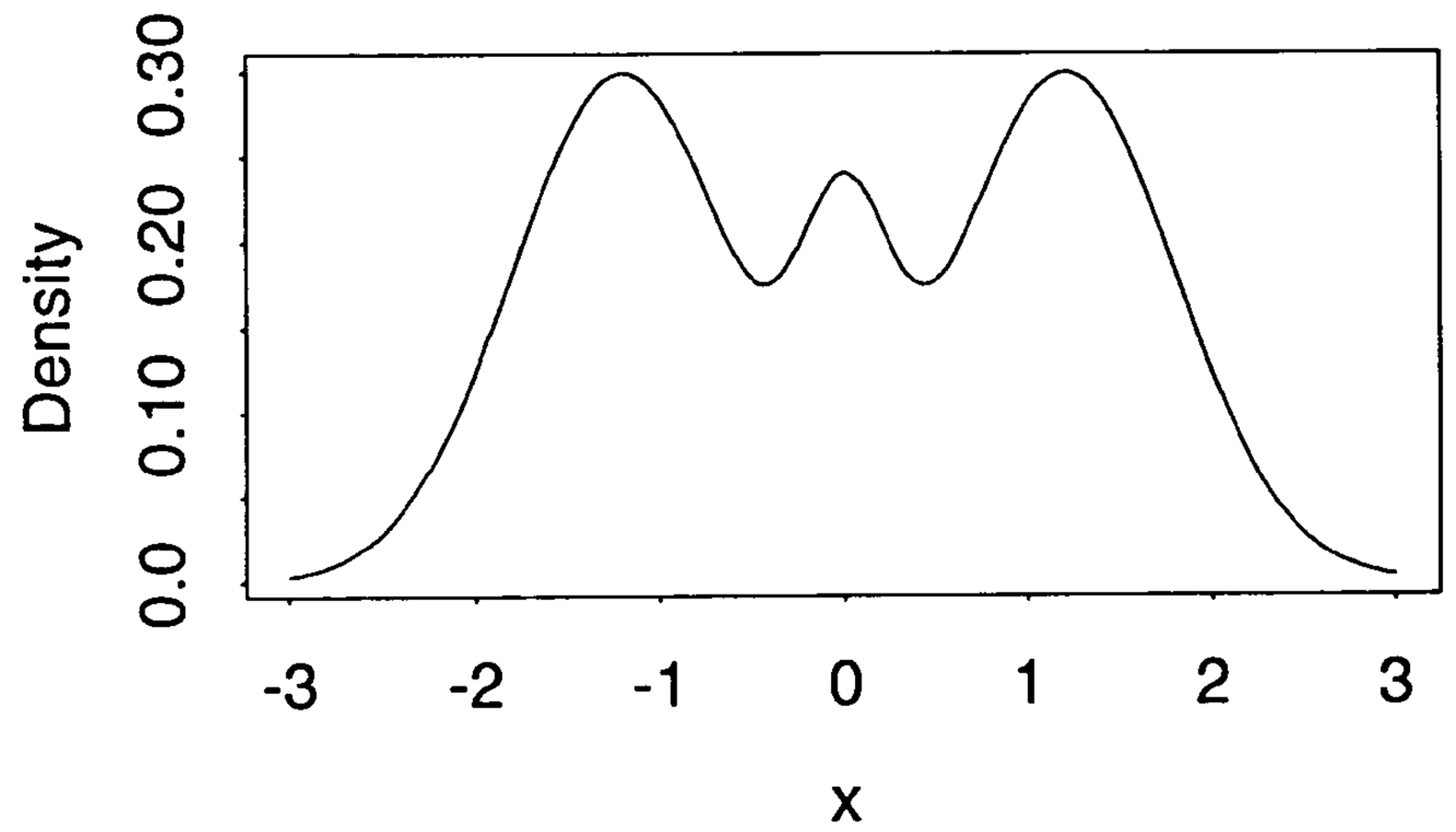


Figure 4.7: Skewed bimodal



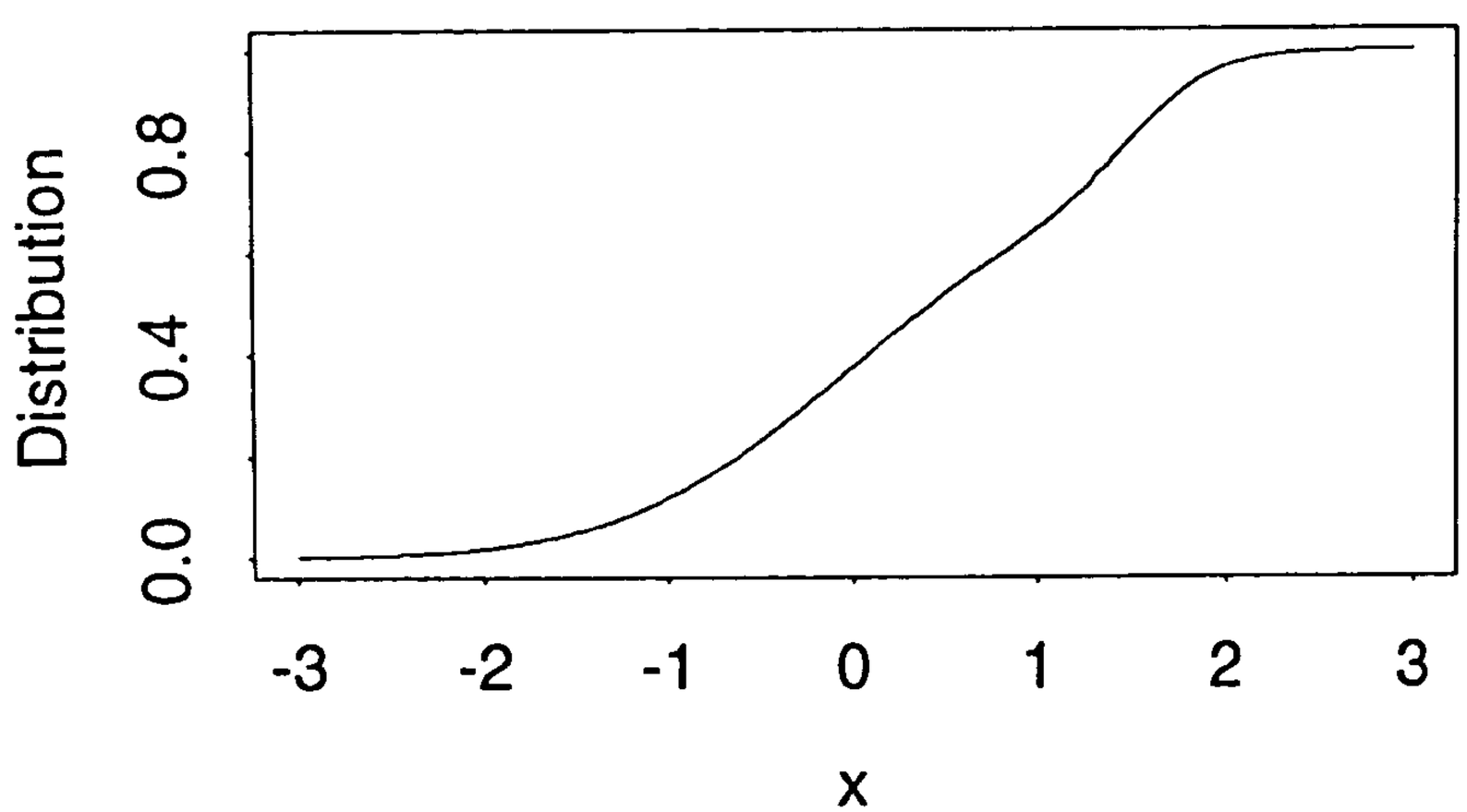


Figure 4.8: Trimodal

p	RE (%) of $F_{e,n}(q_p\|x)$	RE (%) of $F_{s,n}(q_p\|x)$
$p = 0.5$	0.354	0.636
$p = 0.9$	0.985	0.712
$p = 0.1$	0.931	0.479

Table 4.3: RE with Skewed unimodal distribution

p	RE (%) of $F_{e,n}(q_p\|x)$	RE (%) of $F_{s,n}(q_p\|x)$
$p = 0.5$	0.506	0.859
$p = 0.9$	0.967	0.621
$p = 0.1$	0.865	0.848

Table 4.4: RE with Kurtotic unimodal distribution

p	RE (%) of $F_{e,n}(q_p\|x)$	RE (%) of $F_{s,n}(q_p\|x)$
$p = 0.5$	0.136	0.238
$p = 0.9$	0.763	0.984
$p = 0.1$	0.713	0.961

Table 4.5: RE with Outlier distribution

p	RE (%) of $F_{e,n}(q_p\|x)$	RE (%) of $F_{s,n}(q_p\|x)$
$p = 0.5$	0.218	0.275
$p = 0.9$	0.604	0.941
$p = 0.1$	0.983	0.899

Table 4.6: RE with Bimodal distribution

p	RE (%) of $F_{e,n}(q_p\|x)$	RE (%) of $F_{s,n}(q_p\|x)$
$p = 0.5$	0.006	0.006
$p = 0.9$	0.905	0.987
$p = 0.1$	0.702	0.905

Table 4.7: RE with Separated bimodal distribution

p	RE (%) of $F_{e,n}(q_p\|x)$	RE (%) of $F_{s,n}(q_p\|x)$
$p = 0.5$	0.285	0.916
$p = 0.9$	1.000	0.527
$p = 0.1$	1.000	0.117

Table 4.8: RE with Skewed bimodal distribution

p	RE (%) of $F_{e,n}(q_p\|x)$	RE (%) of $F_{s,n}(q_p\|x)$
$p = 0.5$	0.369	0.580
$p = 0.9$	0.973	0.746
$p = 0.1$	0.607	1.000

Table 4.9: RE with Trimodal distribution

From Table 4.2-4.9 in general the double-kernel with double-bandwidth is always best among three estimators while the performance of the other two estimators is a little bit more complex.

First, when the values of y are near median ($p = 0.5$), double-kernel with one bandwidth always does better than single-kernel method; secondly, for extreme values of y, and with Gaussian DF, Outlier DF and Separated bimodal DF, double-kernel with one bandwidth are better than single-kernel method, however, for Skewed DF (unimodal or bimodal) and Kurtotic DF, single-kernel perform better particularly.

But the most interesting conclusion is that two bandwidths are much better (usually) near median, but less so in extremes. This should translate also to quantile estimation, and supports what we said about h_2 in double-kernel quantile estimation.

4.4 Bandwidth Selection

The proposed estimators 4.1-4.3 of $F(y|x)$ are based on single (double) bandwidth. The asymptotic optimal bandwidths h_e and h_s for $F_{e,n}(y|x)$ and $F_{s,n}$ in terms of Integrated Mean Square Error (MISE) are, respectively:

$$h_e(y) = \left(\frac{R(K)\int\int g^{-1}(x)F(y|x)(1-F(y|x))dxdy}{\mu_2^2(K)\int\int\{F^{2,0}(y|x)\}^2dxdy}\right)^{1/5} n^{-1/5}$$

$$h_s(y) = \left(\frac{R(K)\int\int g^{-1}(x)F(y|x)(1-F(y|x)dxdy}{\int\int[\mu_2(K)\{F^{2,0}(y|x)\}+\mu_2(W)\{f'(y|x)\}]^2dxdy}\right)^{1/5} n^{-1/5} \quad (4.7)$$

Explicit expressions for bandwidths $h_d(y)$ and $b(y)$ for estimator $F_{d,n}(y|x)$ are not available. Combining theoretical results and simulation, the following approach is suggested for bandwidth selection:

i) Given y, if $\int\int\{f'(y|x)\}^2dxdy$ is very small, taking both $h_s(y)$ and $h_d(y)$ equals $h_e(y)$ from equation (4.7) as

$$\Big(\frac{R(K)}{\mu_2^2(K)}\frac{\int g^{-1}(x)F(y|x)(1-F(y|x))v(x)dx}{\int\{F^{2,0}(y|x)\}^2v(x)dx}\Big)^{1/5}n^{-1/5}$$

Further when $\int\{f'(y|x)\}^2v(x)dx$ is very small, so is $\int\{F^{2,0}(y|x)\}\{f'(y|x)\}v(x)dx$, and since $\Big(\int\{F^{2,0}(y|x)\}\{f'(y|x)\}v(x)dx\Big)^2 \leq \int\{f'(y|x)\}^2v(x)dx\int\{F^{2,0}(y|x)\}^2v(x)dx$, so the mean square error involving in b is dominated by variance term. So b might be selected to zero the leading variance terms of $\int\int AMISE_d(x,y)dxdy$:

$$b=\frac{1}{\alpha(W)}\frac{\int\int\frac{F(y|x)(1-F(y|x))}{g(x)}dxdy}{\int\int\frac{f(y|x)}{g(x)}dxdy}$$

ii) Given y, if $\int\int\{f'(y|x)\}^2dxdy$ isn't small, then h_e and h_s are selected according to equation (4.7) and $h_d=h_s$ while b is selected according to

$$b=\frac{R(W)\alpha(W)}{n}\Big(\frac{2}{\mu_2(W)^2h}\frac{\int\int\frac{f(y|x)}{g(x)}dxdy}{\int\int(f'(y|x))^2dxdy}\Big)^{1/3}$$

from the proof of Theorem 4.3.

Although there are different choices for the integrating weight $v(x,y)$, $v(x,y)=g(x)$ usually brings about most convenience. These ideas have not been followed up further.

Chapter 5

Local Polynomial Kernel Smoothing Quantiles for Semi-Parametric Likelihood-Based Models

5.1 Introduction

In the present chapter *semi-parametric* methods are discussed, these have some attractive features, and are more flexible concerning the assumptions about the model and the distribution than parametric ones. In essence, these methods involve a parametric transformation of the original response variables, and the quantile curves are given in nonparametric forms. The key issue for this technique is the selection of an appropriate transformation for a given distribution. For example, Box-Cox power transformation (Cole, 1988, Cole and Green, 1992,

Van't Hof *et al*, 1988), Johnson transformation (Thompson and Theron, 1990) and logarithmic transformation (Royston, 1991) all to the normal distribution. Methods of estimating unknown parameters, however, can be parametric or non-parametric. Usually it is required to smooth the estimators to match the quantile values which change smoothly over covariates, particularly in longitudinal data. These estimators are usually maximum likelihood estimators subject to smoothness constraints. If nonparametric estimating methods are adopted, the whole methods are *semi-parametric* ones. The two typical nonparametric smoothing are roughness penalty and kernel estimation; Cole and Green gave a good approach by roughness penalty.

Generally, the transformation for a specified distribution involves several parameters, for example, the Box-Cox transformation of Cole and Green has three parameters while Johnson transformation of Thompson and Theron has four parameters.

Example 1. Suppose that $y^{\lambda}|x \sim N(\mu, \sigma^2)$, λ is the transformation parameter while μ, σ are the distribution parameters. These parameters are functions of the covariate variable x which will be denoted jointly by $\theta(x)$, and are assumed to be continuous. Then the *pth* conditional quantile of y for $X = x$ based on this transformation is $q_p(x, \theta(x)) = z_p(x)^{1/\lambda(x)}$ with $z_p(x) = \sigma(x)\Phi^{-1}(p) + \mu(x)$ (where $\Phi^{-1}(p)$ is the *pth* quantile of standard normal). Alternatively, suppose that $\frac{(\frac{y}{\mu})^{\lambda}-1}{\lambda\sigma}|x \sim N(0,1)$, then there are three transformation parameters and no distribution parameter, also $q_p(x;\theta(x)) = \mu\Big(\lambda\sigma z_p(x) + 1\Big)^{1/\lambda}$, with $z_p(x) = \Phi^{-1}(p)$ here. Particularly, when the power parameter $\lambda = 0$, this will give a *log-transformation*, i.e. $\frac{log\{\frac{y}{\mu}\}}{\sigma} \sim N(0,1)$. Here μ, λ, σ are also functions of x.

Example 2. Suppose $y^{\lambda}|x \sim \Gamma(\sigma, 1/\sigma)$, with transformation parameter λ, and

distribution parameter σ respectively. Both λ and σ are functions of x.

Example 3. S_u transformation: $y|x = \epsilon + \lambda sinh(\frac{Z-\mu}{\sigma})$ with $Z \sim N(0,1)$. There are four transformation parameters ϵ, λ, μ and σ.

Generally, for a random vector (X, Y) where X and Y are the covariate and response, let $q_p(x)$ be the conditional p-quantile of Y given $X = x$ and $\theta = (\theta_1, \theta_2, ..., \theta_m)^T$ (usually $1 < m \leq 4$) be the vector of distribution and transformation parameters. The transformation $Z_\theta = Q(\theta, Y)$ from Y to Z is supposed to have known conditional distribution $f(z)$ given $X = x$ for which any conditional p-quantile z_p $(0 < p < 1)$ can explicitly be obtained . It is reasonable and practical to assume that the value of all the parameters θ depend on the covariate x. Then the original *pth* conditional quantile $q_p(x)$ of Y given $X = x$ is given by

$$q_p(x, \theta(x)) = Q^{-1}(\theta(x), z_p). \tag{5.1}$$

The assumption that $Q(\theta, .)$ is monotone increasing in p is a natural guarantee that the quantiles do not cross: as long as the transformation $Q(\theta(x), .)$ is one-one monotone mapping, the conditional quantile defined by equation (5.1) still keeps this non-decreasing property which is an appealing one of semi-parametric smoothing quantiles. That is, if $Q(\theta(x); u)$ is differentiable with respect to u, and since $\frac{d}{dz_p}Q^{-1}(\theta(x), z_p) \geq 0$ and $Q(\theta(x); u)$ is monotone increasing, then

$$\frac{d}{dp}q_p(x; \theta(x)) = \frac{d}{dz_p}\Big(Q^{-1}(\theta(x), z_p)\Big)\frac{d}{dp}(z_p) \geq 0.$$

Given n independent observations $\{y_i\}$ and corresponding covariate values $\{x_i\}$, it is convenient to use maximum likelihood method of estimation to estimate θ; however, this would result in an unsatisfactory estimate due to overfitting. Some techniques to quantify local variation are adopted. For example, Cole and Green used roughness penalty by cubic spline, which requires m smoothing parameters. Replacing roughness penalty, kernel-weighting techniques of local constant fitting

and local linear fitting, with a single kernel and a single smoothing parameter are investigated. So, if $\hat{\theta}(x)$ is the smoothing estimator of $\theta(x)$, then

$$\hat{q}_p(x) = Q^{-1}(\hat{\theta}(x), z_p). \tag{5.2}$$

5.2 Locally Kernel-Weighted Maximum Likelihood

5.2.1 The Model

The density function of $Z = Q(\theta, Y)$ is proportional to $f(z)$ and to change back to Y

$$f(z)\frac{dz}{dy} = f(Q(\theta, y))Q_y(\theta, y)$$

is needed where $Q_y(\theta, y)$ stands for the derivative of $Q(\theta, y)$ with respect to y.

Moreover, the log density function of Y in terms of parameter vector $\theta = (\theta_1, ..., \theta_m)^T$ is

$$L(\theta; Y) = log\{f(Q(\theta(x), Y))Q_y(\theta(x), Y)\} \tag{5.3}$$

while the corresponding log-likelihood function based on n observations is

$$l_\theta = \sum_j log\{f(Q(\theta(x), Y_i))Q_y(\theta(x), Y_i)\}. \tag{5.4}$$

Consider a simple example where $Z = \frac{Y-\mu}{\sigma} \sim N(0,1)$, then $f(z) \propto e^{-1/2z^2}$, and the log density function of Y is $\left(-\frac{1}{2}Z^2 + log(\frac{dz}{dy})\right)$ or log-likelihood $\sum_i \left(-\frac{1}{2}z_i^2 + log(\frac{dz_i}{dy})\right)$. Thus, for Box-Cox transformation of Cole and Green, $z = \frac{(y/\mu)^\lambda - 1}{\sigma\lambda}$, $\frac{d}{dy}z = \frac{d}{dy}Q(\theta, y) = \frac{1}{\sigma y}(y/\mu)^\lambda$, the log density function of Y in terms of λ, μ, σ is

$\lambda log\frac{y}{\mu} - log\sigma - \frac{1}{2}z^2$, and the log-likelihood function for $(\lambda, \mu, \sigma)^T$ is

$$\sum_j \left(\lambda log\frac{y_j}{\mu} - log\sigma - 1/2z_j^2\right), \quad z_j = \frac{(y_j/\mu)^\lambda - 1}{\sigma\lambda}, \quad j = 1, 2, ..., p.$$

To introduce the idea of local polynomial fitting $\theta(.) = (\theta_1(.), ..., \theta_m(.))^T$, let

$$\theta(x) + \theta^{(1)}(x)(X_i - x) + ... + \frac{\theta^{(q)}(x)}{q!}(X_i - x)^q$$
$$\equiv \beta_0(m) + \beta_1(m)(X_i - x) + ... + \beta_q(m)(X_i - x)^q.$$

for X_i in a neighborhood of x, where the $\theta^{(1)}(x), ..., \theta^{(q)}(x)$ are the derivative functions of $\theta(x)$, all of them are m-dimensional vector; $\beta_0(m), ..., \beta_q(m)$ are also the vector with $\beta_t(m) = (\beta_{1t}, \beta_{2t}, ..., \beta_{mt})^T$ $(t = 0, 1, ..., q)$, and for a symmetric kernel K, observations on X_i in the log-likelihood are weighted by $K_h(X_i - x)$.

Let $\hat{\beta}_t(m) = (\hat{\beta}_{1t}, ..., \hat{\beta}_{mt})^T$ and $\hat{\theta}(x; q, h) = (\hat{\theta}_1(x; q, h), ..., \hat{\theta}_m(x; q, h))^T$ be m-dimensional vector estimators of $\beta_t(m)$ $t = 0, 1, ..., q$ and $\theta(x; q, h)$.

Then local polynomial kernel estimator of each $\theta(x)$ is given by $\hat{\theta}(x; q, h) = \hat{\beta}_0(m)$ and $\hat{\beta}_t(m)$ $t = 0, 1, ..., q$ maximizes

$$l(\beta_0(m), \beta_1(m), \quad ... \ , \beta_q(m)) =$$
$$\sum_i log\{f(Q(\beta_0(m) + \beta_1(m)(X_i \quad - x) + ..+ \beta_q(m)(X_i - x)^q, Y_i))$$
$$\times Q_y(\beta_0(m) + \beta_1(m)(X_i - x) + ... + \beta_q(m)(X_i - x)^q, Y_i)\}K_h(X_i - x) \quad (5.5)$$

When $m = 1$ and $q = 0$, this model gives local constant kernel fitting single parameter considered by Staniswalis (1989), and if $m = 1$ corresponds to the case considered by Fan, Heckman and Wand (1995). The basic idea of local likelihood estimate like this goes back to Tibshirani and Hastie (1987).

Local polynomial fitting provides consistent estimates of higher-order derivatives of θ through the coefficients of higher-order terms in the polynomial fit. Define

an estimator of $\theta^{(r)}(x)$ for $r = 0, 1, ..., q$ to be

$$\hat{\theta}_r(x; p, h) = r!\hat{\beta}_r(m).$$

The main concern here is $\hat{\theta}(x; p, h)$, and example of this is the Box-Cox transformation of Cole model.

Let $f_Y(y|x)$ and $F_Y(y|x)$ be the conditional density and distribution functions in Y-space, and $f_Z(z; \theta|x)$ and $F_Z(z; \theta|x)$ be the corresponding conditional density and distribution in Z-space, i.e. $Z = \frac{(Y/\mu)^\lambda - 1}{\lambda\sigma}$ and the p-quantile z_p is determined by $f_Z(z; \theta|x)$ and $F_Z(z; \theta|x)$. Clearly, $q_p(x, \theta(x)) = \mu(\lambda\sigma z_p + 1)^{1/\lambda} = q_p(x)$.

5.2.2 Asymptotic MSE for $\hat{q}_p(x)$

Since

$$\hat{q}_p(x) - q_p(x; \theta(x)) \approx \Big(\frac{\partial Q^{-1}(\theta, z_p)}{\partial\theta(x)}\Big)^T \Big(\hat{\theta}(x) - \theta(x)\Big),$$

that is, $\hat{q}_p(x) - q_p(x; \theta(x))$ is approximately a linear combination of the components $\hat{\theta}_t(x) - \theta_t(x)$, $t = 1, 2, ..., m$ of vector $\hat{\theta}(x) - \theta(x)$, and it is asymptotically normal if $\hat{\theta}(x) - \theta(x)$ is asymptotically normal under the same conditions for which $\hat{\theta}(x) - \theta(x)$ has normal distribution.

Thus, if

$$\sqrt{nh}\Big(\hat{\theta}(x) - \theta(x)\Big) \sim N\Big(Bias(x, K, h, g), Var(x, K, h, g)\Big),$$

then

$$\sqrt{nh}\Big(\hat{q}_p(x) - q_p(x)\Big)$$

$$\sim N\Big((\frac{\partial Q^{-1}(\theta, z_p)}{\partial\theta(x)})^T Bias(x, K, h, g), (\frac{\partial Q^{-1}(\theta, z_p)}{\partial\theta(x)})^T Var(x, K, h, g)\frac{\partial Q^{-1}(\theta, z_p)}{\partial\theta(x)}\Big).$$

and the mean square error of $\hat{q}_p(x)$ is

$$\begin{aligned} MSE(\hat{q}_p(x)) &= \{(\frac{\partial Q^{-1}(\theta, z_p)}{\partial \theta(x)})^T Bias(x, K, h, g)\}^2 \\ &+ \frac{1}{nh}(\frac{\partial Q^{-1}(\theta, z_p)}{\partial \theta(x)})^T Var(x, K, h, g) \frac{\partial Q^{-1}(\theta, z_p)}{\partial \theta(x)} \end{aligned} \tag{5.6}$$

5.3 Asymptotic Theorems

Suppose that the covariate variable X has density $g(x)$ with support $(0, 1)$, though usually in most applications represent age. In this section asymptotic properties of the estimates are developed in the interior of $(0, 1)$ and near boundary. Let a symmetric kernel K has limited support [-1,1], and if $x = 0 + \alpha h$ or $x = 1 - \alpha h$ for some $0 \leq \alpha < 1$, then x is near the boundary and it is called boundary point.

5.3.1 Local q-Order Polynomial Fitting

The main result here is given in the following general theorem.

Theorem 5.1. Suppose that

i) The density $g(x)$ of covariate has continous first-order derivative, and kernel K is a symmetric density with support[-1, 1].

ii) Denote simply the log-likelihood function (5.5) as $l(\beta, q; Y)$, and $l(\beta, q; Y)$ has continuous third-order derivative for any number n of sample.

iii) Fisher-Information matrix $I(\theta) = -E\frac{\partial^2 L(\theta;Y)}{\partial \theta^2}$ exists and it is positive-definite.

iv) Let $a_n = \frac{1}{\sqrt{nh}}$, and asymptotically $a_n^2 E\frac{\partial l(\beta,q;Y)}{\partial \beta}$ and $a_n^2 Cov\{\frac{\partial l(\beta,q;Y)}{\partial \beta}\}$ exist and are bounded.

Then under local q-order polynomial fitting when $nh \to +\infty$ and $h \to 0$,

$$\sqrt{nh}\Big[\Big(\hat{\beta}_0(m) - \beta_0(m), ..., \hat{\beta}_q(m) - \beta_q(m)\Big)^T - a_n^2 A(x, g, K, \beta(m))^{-1} E\frac{\partial l(\beta, q; Y)}{\partial \beta}\Big]$$

has an asymptotic distribution given by

$$N\Big(0_q(m), A(x, g, K, \beta(m))^{-1} \times a_n^2 Cov\{\frac{\partial l(\beta, q; Y)}{\partial \beta}\} A(x, g, K, \beta(m))^{-1}\Big).$$

Here $A(x, g, K, \theta)$ is positive-definite matrix which is composed of $(q+1)^2$ m-order sub-matrices such that $(j, t)th$ sub-matrix is

$$A(x, g, K, \theta)_{jt} = (-1)^{j+t+1} h^{j+t} I(\theta) g(x) \int z^{j+t} K(z) dz,$$

$$((j, t) = (0, 0), (0, 1), (1, 1), ..., (q+1, q+1)).$$

When design density $g(x)$ has a boundary support, say (0,1), and $x = x_n$ is a left boundary point, the above conclusions still hold with integrals $\int z^{j+t} K(z) g(x + hz) dz$ and $\int K(z) dz$ replaced by $\int_D z^{j+t} K(z) g(x + hz) dz$ and $\int_D K(z) dz$ respectively , where $D = \{z : x - hz \in (0, 1)\} \cap [-1, 1]$.

Particularly, under local linear fitting,

$$A(x, g, K, \beta) = -\begin{pmatrix} I(\theta) g(x) & 0 \\ 0 & h^2 I(\theta) \mu_2(K) g(x) \end{pmatrix}.$$

and under local constant fitting,

$$A(x, g, K, \beta(m))) = -I(\theta) g(x).$$

Corollary 5.1. Under the conditions of Theorem 5.1, and if

(1) with local linear fitting, then

$$\sqrt{nh}\Big[\Big(\hat{\theta}(x;1,h)-\theta(x)\Big)^T+(I(\theta)g(x))^{-1}a_n^2E\frac{\partial l(\beta,1;Y)}{\partial\beta}\Big]$$

$$\sim N\Big(0,(I(\theta)g(x))^{-1}a_n^2Cov\frac{\partial l(\beta,1;Y)}{\partial\beta}(I(\theta)g(x))^{-1}\Big);$$

(2) with local constant fitting, then

$$\sqrt{nh}\Big[\Big(\hat{\theta}(x;0,h)-\theta(x)\Big)^T+(I(\theta)g(x))^{-1}a_n^2E\frac{\partial l(\beta,0;Y)}{\partial\beta}\Big]$$

$$\sim N\Big(0,(I(\theta)g(x))^{-1}a_n^2Cov\frac{\partial l(\beta,0;Y)}{\partial\beta}(I(\theta)g(x))^{-1}\Big).$$

Proof: Since $(\hat{\beta}_0(m),...,\hat{\beta}_q(m))^T$ maximizes equation (5.5), and from a Taylor expansion of log-likelihood $l(\beta,q;Y)$,

$$\begin{aligned}0_q(m) &= a_n\frac{\partial l(\beta,q;Y)}{\partial\hat{\beta}}=a_n\frac{\partial l(\beta,q;Y)}{\partial\beta}\\ &+ a_n^2\frac{\partial^2 l(\beta,q;Y)}{\partial\beta^2}\times\sqrt{nh}(\hat{\beta}(m)-\beta(m))\\ &+ \sqrt{nh}(\hat{\beta}(m)-\beta(m))^T\times a_n^3\frac{\partial^3 l(\beta,q;Y)}{\partial\beta(*)^3}\times\sqrt{nh}(\hat{\beta}(m)-\beta(m))\end{aligned}\quad(5.7)$$

where

$$\begin{aligned}0_q(m) &= (0,...,0)^T\\ \frac{\partial l(\beta,q;Y)}{\partial\hat{\beta}} &= (\frac{\partial l(\beta,q;Y)}{\partial\hat{\beta}_0(m)},...,\frac{\partial l(\beta,q;Y)}{\partial\hat{\beta}_q(m)})^T\\ \frac{\partial l(\beta,q;Y)}{\partial\beta} &= (\frac{\partial l(\beta,q;Y)}{\partial\beta_0(m)},...,\frac{\partial l(\beta,q;Y)}{\partial\beta_q(m)})^T\\ (\hat{\beta}(m)-\beta(m))^T &= (\hat{\beta}_0(m)-\beta_0(m),...,\hat{\beta}_q(m)-\beta_q(m))^T\end{aligned}$$

and $\frac{\partial^2 l(\beta,q;Y)}{\partial\beta^2}$ is the matrix of the second derivatives of $l(\beta,q;Y)$ with respect to β, also $\frac{\partial^3 l(\beta,q;Y)}{\partial\beta(*)^3}$ is the third derivative matrix with vector $\beta(*)$ between $\hat{\beta}(m)$ and $\beta(m)$.

Let matrix

$$A_n = a_n^2 \frac{\partial^2 l(\beta, q; Y)}{\partial \beta^2}$$

and for $i = 1, ..., n$, and $j, t = 0, 1, ..., q$, define

$$M\ (\beta_0(m); Y_i) = log\{f(Q(\beta_0(m); Y_i))Q_y(\beta_0(m); Y_i)\} \tag{5.8}$$

$$\begin{aligned} M\ (\beta_0(m), \beta_1(m), ..., \beta_q(m); (X_i - x), Y_i) \\ = \quad log\{f(Q(\beta_0 + \beta_1(X_i - x) + ... + \beta_q(X_i - x)^q, Y_i)) \\ \times \quad Q_y(\beta_0 + \beta_1(X_i - x) + ... + \beta_q(X_i - x)^q, Y_i)\} \end{aligned}$$

then

$$\begin{aligned} \frac{\partial}{\partial \beta_t(m)} \quad & M(\beta_0(m), \beta_1(m), ..., \beta_q(m); (X_i - x), Y_i) \\ = \quad & (X_i - x)^t \frac{d}{dB} M(B; Y_i)|_{B=\beta_0+\beta_1(X_i-x)+...+\beta_q(X_i-x)^q} \qquad (5.9) \\ \frac{\partial^2}{\partial \beta_j(m) \partial \beta_t(m)} \quad & M(\beta_0(m), \beta_1(m), ..., \beta_q(m); (X_i - x), Y_i) \\ = \quad & (X_i - x)^{t+l} \frac{d^2}{dB^2} M(B; Y_i)|_{B=\beta_0+\beta_1(X_i-x)+...+\beta_q(X_i-x)^q} \end{aligned}$$

Clearly, the log-likelihood with general local polynomial kernel fitting now is

$$l(\beta_0(m), ..., \beta_q(m)) = \sum_{i=1}^{n} M(\beta_0(m), \beta_1(m), ..., \beta_q(m); (X_i - x), Y_i) K(\frac{X_i - x}{h}) \tag{5.10}$$

then from

$$(A_n)_{jt} = E(A_n)_{jt} + O_p[\{var(A_n)_{jt}\}^{1/2}]$$

$$\begin{aligned} E(A_n)_{jt} \quad = \quad & a_n^2 \sum_{i=1}^{n} E \frac{d^2}{d\beta_j(m) d\beta_t(m)} M(\beta_0(m); Y_i)(X_i - x)^{j+t} K(\frac{X_i - x}{h}) \\ \approx \quad & (-1)^{j+t+1} h^{j+t} \int z^{j+t} E_Y[M(\beta_0 + \beta_1 zh + ... + \beta_q z^q h^q; Y_1)] K(z) g(x + hz) dz \\ = \quad & (-1)^{j+t+1} h^{j+t} I(\beta_0(m)) g(x) \int z^{j+t} K(z) dz + o(h) \end{aligned}$$

so

$$E(A_n) = A(x, g, K, \beta) + o(h).$$

Similar argument shows that

$$var(A_n)_{jt} = O\{(nh)^{-1}\},$$

as

$$var(A_n)_{jt} = a_n^4 \sum_{i=1}^{n} var \frac{d^2}{d\beta_j(m)d\beta_t(m)} M(\beta_0(m); Y_i)(X_i - x)^{2(j+t)} K^2(\frac{X_i - x}{h}).$$

The matrix of third derivatives in the last expansion (5.7) is of order $O_P\{(nh)^{-1/2}\}$, then

$$\sqrt{nh}(\hat{\beta}_0(m) - \beta_0(m), ..., \hat{\beta}_q(m) - \beta_q(m))^T$$

has the same asymptotic distribution as

$$A(x, g, K, \beta(m))^{-1} \times a_n(\frac{\partial l}{\partial \beta_0(m)}, ..., \frac{\partial l}{\partial \beta_q(m)})^T.$$

To derive the distribution of

$$A(x, g, K, \beta(m))^{-1/2}\Big[a_n(\frac{\partial l}{\partial \beta_0(m)}, ..., \frac{\partial l}{\partial \beta_q(m)})^T - a_n E(\frac{\partial l}{\partial \beta_0(m)}, ..., \frac{\partial l}{\partial \beta_q(m)})^T\Big],$$

invoke Cramer-Wold device (to prove the asymptotic normality of a random vector sequence V_n, consider any linear combination with unit vector u: $u^T\{var(V_n)\}^{-1/2}(V_n - EV_n)$, if latter asymptotic $N(0,1)$ by checking Lyapounov's conditions, then V_n is asymptotic normal) it can easily be verified this has an asymptotic normal distribution $N(0, I_{m(q+1)})$. Eventually, Theorem 5.1 follows. Under local linear fitting, the conclusion is derived from

$$\int z^{j+t} K(z)dz = \begin{cases} 0 & \text{if } j+t \text{ is odd} \\ \mu_{j+t}(K) & \text{otherwise} \end{cases}$$

so is $E(A_n)_{0,1} = E(A_n)_{1,0} = 0$.

5.3.2 Simultaneous Fitting Mean Function and Variance Function

Higher-order polynomial fitting involves complicated matrix computation as shown in Theorem 5.1. This section is mainly concerned with local constant and linear fitting. And by the way of application, I first consider normal-based smooth mean and variance fitting only.

As in the previous chapter consider a random vector (X, Y) and let $m(x) = E\{Y|X = x\}$ and $v(x) = Var\{Y|X = x\}$ be the mean and variance of Y when $X = x$. The kernel-weighting estimation problem of $m(x)$ and $v(x)$ is discussed widely in the literature (cf. Fan & Gijbels, 1995, Ruppert, Wand, Holst & Holsjer, 1995).

Suppose $Z = \frac{Y-\mu(x)}{\sqrt{v(x)}} \sim N(0,1)$ for $X = x$. Then the log-likelihood for local constant fitting $(m(x), v(x))^T$ is

$$l(m(x), v(x)) = -\sum_{i=1}^{n} \frac{(Y_i - m)^2}{v} K(\frac{X_i - x}{h}) - \sum_{i=1}^{n} (logv) K(\frac{X_i - x}{h}) \qquad (5.11)$$

Then the resulting estimators are simply

$$\hat{m}(x) = \sum_i w(x, K) Y_i, \quad \hat{v}(x) = \sum_i w(x, K)[Y_i - \hat{m}(x)]^2,$$

where $w(x, K) = \frac{K(\frac{X_i - x}{h})}{\sum_i K(\frac{X_i - x}{h})}$.

Theorem 5.2. Under the regularity conditions and with local constant fitting,

$$\begin{aligned} \sqrt{nh}\Big[\Big(\hat{m}(x) - m(x) \quad &, \quad \hat{v}(x) - v(x)\Big)^T \\ -h^2\mu_2(K)\Big(\frac{1}{2}m''(x) + \frac{m'(x)g'(x)}{g(x)} \quad &, \quad \frac{1}{2}v''(x) + m'(x)^2 + v'(x)\frac{g'(x)}{g(x)}\Big)^T\Big] \\ &\to \quad N(0, COV_0) \end{aligned}$$

where the covariance matrix $COV_0 = \frac{R(K)}{g(x)}C$ and matrix C has diagonal elements $v(x)$, $E\{Y - m(x)\}^4 - v^2(x)$ and off diagonal element $E\{Y - m(x)\}^3$.

Proof: Let $a_n = (nh)^{-1/2}$, asymptotically from Corollary 5.1

$$\sqrt{nh}\Big[\hat{m}(x) - m(x), \hat{v}(x) - v)^T - \{I(m(x), v(x))g(x)\}^{-1}a_n^2 E\frac{\partial l(m(x), v(x))}{\partial(m(x), v(x))^T}\Big]$$

$$\approx N\Big(0, \{I(m(x), v(x))g(x)\}^{-1}a_n^2 Cov\{\frac{\partial l(m(x), v(x))}{\partial(m(x), v(x))^T}\}\{I(m(x), v(x))g(x)\}^{-1}\Big).$$

It is easy to show that

$$-I(m(x), v(x))g(x) = diag\{\frac{2}{v(x)}, \frac{1}{v(x)^2}\}g(x) + o(h).$$

Now to calculate $a_n^2 E\frac{\partial l(m(x),v(x))}{\partial(m(x),v(x))^T}$, consider

$$\begin{aligned} E(Y_i - m) &= m(X_i) - m(x) \\ &= m'(x)(X_i - x) + \frac{1}{2}m''(x)(X_i - x)^2 + o(h^2) \\ E(Y_i - m)^2 &= E(Y_i - m(X_i))^2 + (m(X_i) - m(x))^2 \\ = v(x) + v'(x)(X_i - x) &+ \frac{1}{2}v''(x)(X_i - x)^2 + m'(x)^2(X_i - x)^2 + o(h^2) \end{aligned}$$

then

$$\begin{aligned} a_n^2 E\frac{\partial l(m(x), v(x))}{\partial(m(x), v(x))^T} &= o(h^2) - h^2\mu_2(K)g(x) \\ &\times(\frac{2}{v(x)}\mu'(x)\frac{g'}{g} + \frac{\mu''(x)}{v(x)} \quad , \quad \frac{v'(x)}{v(x)^2}\frac{g'}{g} + \frac{1}{2}\frac{v''(x)}{v(x)^2} + \frac{\mu'(x)^2}{v^2})^T. \end{aligned}$$

Continuous to calculate $a_n^2 Cov\{\frac{\partial l(m(x),v(x))}{\partial(m(x),v(x))^T}\}$, consider

$$\begin{aligned} var\{a_n\sum_{i=1}^{n}\frac{2(Y_i - m)}{v}K(\frac{X_i - x}{h})\} &= h^{-1}var\{\frac{2(Y_1 - m)}{v}K(\frac{X_1 - x}{h})\} \\ &= h^{-1}\frac{4}{v(x)^2}\int v(u)K^2(\frac{u - x}{h})g(u)du \\ &= 4v(x)^{-1}R(K)g(x) + o(h) \end{aligned}$$

Similarly,

$$\begin{aligned} var\{-a_n \sum_{i=1}^{n} \frac{(Y_i - m)^2}{v^2} K(\frac{X_i - x}{h})\} &= h^{-1} var\{\frac{(Y_1 - m)^2}{v^2} K(\frac{X_1 - x}{h})\} \\ &= \frac{1}{v(x)^4} R(K) g(x) (E\{Y - m(x)\}^4 - v(x)^2) \\ &+ o(h) \end{aligned}$$

and

$$\begin{aligned} a_n^2 \sum_{i=1}^{n} \sum_{j=1}^{n} K(\frac{X_i - x}{h}) K(\frac{X_j - x}{h}) &\times Cov\{-2\frac{(Y_i - m)}{v}, \frac{1}{v} - \frac{(Y_j - m)^2}{v^2}\} \\ &= \frac{2R(K)}{v(x)^3} g(x) E\{(Y - m(x))^3 | X = x\} \\ &+ o(h) \end{aligned}$$

and hence Theorem 5.2.

Theorem 5.3. The log-likelihood fit of $(m(x), v(x))^T$ by local linear kernel maximizes

$$\begin{aligned} l(m(x), v(x); m_1(x), v_1(x)) &= -\sum_{i=1}^{n} \frac{(Y_i - m - m_1(X_i - x))^2}{v + v_1(X_i - x)} K(\frac{X_i - x}{h}) \\ &- \sum_{i=1}^{n} (log\{v + v_1(X_i - x)\}) K(\frac{X_i - x}{h}); \end{aligned}$$

then, under the regularity conditions,

$$\begin{aligned} \sqrt{nh}\Big[\Big(\hat{m}(x) - m(x), \hat{v}(x) - v(x)\Big)^T &- \frac{1}{2} h^2 \mu_2(K) \Big(m''(x), v''(x)\Big)^T\Big] \\ &\to N(0, COV_1) \end{aligned}$$

where $COV_1 = \frac{R(K)}{g(x)} C$ and matrix C has diagonal elements $v(x)$ and $E\{Y - m(x)\}^4 - v(x)^2$ and off-diagonal element $E\{Y - m(x)\}^3$.

Proof: Let $a_n = (nh)^{-1/2}$, asymptotically from Corollary 5.1

$$\sqrt{nh}\Big[\hat{m}(x) - m(x), \hat{v}(x) - v)^T - \{I(m(x), v(x)) g(x)\}^{-1} a_n^2 E \frac{\partial l(m(x), v(x); m_1(x), v_1(x))}{\partial (m(x), v(x))^T}\Big]$$

$$\approx$$

$$N\Big(0, \{I(\mu(x), v(x))g(x)\}^{-1} a_n^2 Cov\{\frac{\partial l(m(x), v(x); m_1(x), v_1(x))}{\partial (m(x), v(x))^T}\}\{I(m(x), v(x))g(x)\}^{-1}\Big).$$

Now to calculate $a_n^2 E\frac{\partial l(m(x),v(x);m_1(x),v_1(x))}{\partial (m(x),v(x))^T}$ and $a_n^2 Cov\frac{\partial l(m(x),v(x);m_1(x),v_1(x))}{\partial (m(x),v(x))^T}$. Note that for bounded functions a, b and an integer c,

$$\int \frac{t^c K(t)}{a+bht} dt = \frac{1}{a}\int t^c K(t)dt + O(h).$$

From

$$E\{Y - m - m_1(X-x)\} = (m'(x) - m_1(x))(X-x) + \frac{1}{2}m''(x)(X-x)^2 + O(h^2),$$

and kernel K is symmetric,

$$\begin{aligned} a_n^2 E\{\sum_{i=1}^n \frac{Y_i - m - m_1(X_i - x)}{v + v_1(X_i - x)}\}K(\frac{X_i - x}{h}) &= \frac{\frac{1}{2}m''(x)(X-x)^2}{v(x) + v'(x)(X-x)} + O(h^2) \\ a_n^2 var\{\sum_{i=1}^n \frac{Y_i - m - m_1(X_i - x)}{v + v_1(X_i - x)}K(\frac{X_i - x}{h})\} &= 4v(x)^{-1}R(K)g(x) + o(h) \end{aligned}$$

another following results are easily verified

$$\begin{aligned} a_n^2 E\frac{\partial l(m(x), v(x); m_1(x), v_1(x))}{\partial m(x)} &= h^2\mu_2(K)g(x)\frac{m''(x)}{v(x)} + o(h^2) \\ a_n^2 E\frac{\partial l(m(x), v(x); m_1(x), v_1(x))}{\partial v(x)} &= 1/2h^2\mu_2(K)g(x)\frac{v''(x)}{v(x)^2} + o(h^2) \\ a_n^2 Var\frac{\partial l(m(x), v(x); m_1(x), v_1(x))}{\partial m(x)} &= 4v(x)^{-1}R(K)g(x) + o(h) \\ a_n^2 Var\frac{\partial l(m(x), v(x); m_1(x), v_1(x))}{\partial v(x)} &= \frac{1}{v(x)^4}R(K)g(x)(E\{Y - m(x)\}^4 - v(x)^2) + o(h) \\ a_n^2 Cov\{\frac{\partial l(m(x), v(x); m_1(x), v_1(x))}{\partial m(x)} &, \frac{\partial l(m(x), v(x); m_1(x), v_1(x))}{\partial v(x)}\} \\ &= \frac{2R(K)}{v(x)^3}g(x)E\{(Y - m(x))^3 | X = x\} + o(h) \end{aligned}$$

In fact, from

$$E\frac{\partial l(m(x), v(x); m_1(x), v_1(x))}{\partial v(x)} =$$

$$E\Big[\sum_{i=1}^n \frac{(Y_i - m - m_1(X_i - x))^2}{(v + v_1(X_i - x))^2}K(\frac{X_i - x}{h}) - \sum_{i=1}^n \frac{1}{v + v_1(X_i - x)}K(\frac{X_i - x}{h})\Big];$$

and

$$E\{Y - m - m_1(X - x)\}^2 = v(x) + v'(x)(X - x) + \frac{1}{2}v''(x)(X - x)^2$$

$$-(2m_1(x)m'(x) - m'(x)^2 - m_1(x)^2)(X - x)^2 - m_1(x)m''(x)(X - x)^3 + O(h^2),$$

and using $m_1(x) \approx m'(x)$ in last calculation, we have above $a_n^2 E\frac{\partial l(m(x),v(x);m_1(x),v_1(x))}{\partial v(x)}$ and hence Theorem 5.3.

5.4 Kernel Version of LMS Method

The smoothing reference chart method proposed by Cole (1988), known as LMS method, is based on Box-Cox transformation for normal distribution. The transformation is used for three smoothing estimators L(x), M(x) and S(x), which respectively represent estimation of a continuous function of x with power parameter λ, median m of Y and the coefficient of variation (CV) of Y.

For given $X = x$, the transformation

$$Z = \begin{cases} \frac{(Y/\mu)^\lambda - 1}{\lambda\sigma} & \text{if } \lambda \neq 0 \\ \frac{log(Y/\mu)}{\sigma} & \text{otherwise} \end{cases}$$

has a standard normal distribution and the pth percentile of y at x is given by

$$q_p(x) = \mu(x)(1 + \lambda(x)\sigma(x)\Phi^{-1}(p))^{1/\lambda(x)} \tag{5.12}$$

which is estimated by

$$\hat{q}_p(x) = \begin{cases} M(x)(1 + L(x)S(x)\Phi^{-1}(p))^{1/L(x)} & \text{if } L(x) \neq 0 \\ M(x)exp(S(x)\Phi^{-1}(p)) & \text{otherwise} \end{cases}$$

where $\Phi^{-1}(p)$ is the normal equivalent deviate of size p.

5.4.1 Local Constant Fitting LMS

Cole and Green (1992) used roughness penalty for smoothing these three parameters. To obtain a kernel version, the log-likelihood function l_c of n independent observations from (X, Y) in terms of λ, μ and σ and under the Box-Cox transformation of Cole is (apart from the constant)

$$l_c = \sum_i \left(\lambda log \frac{Y_i}{\mu} - log\sigma - 1/2 Z_i(c)^2 \right) \tag{5.13}$$

where $Z(c) \equiv Z$, and c means Z is a local constant version.

Here we obtain estimates of $L(t)$, $M(t)$ and $S(t)$ and hence of the conditional quantiles, by maximizing L_c in terms of kernel K, where

$$L_c = \sum_i \left(\lambda log \frac{Y_i}{\mu} - log\sigma - 1/2 Z_i(c)^2 \right) K((X_i - x)/h). \tag{5.14}$$

Theorem 5.4: Under the regularity conditions and with local constant fitting, and assume that "true model" is normal once λ, μ and σ transformations are done, then

$$\begin{aligned} \sqrt{nh}\Big[& \Big(L(x) - \lambda(x), M(x) - \mu(x), S(x) - \sigma(x) \Big)^T \\ & - \frac{1}{2} h^2 \mu_2(K) I^{-1}(\lambda, \mu, \sigma) \Big(a(x), b(x), 0 \Big)^T \Big] \\ & \rightarrow N\Big(0, \frac{R(K)}{g(x)} I^{-1}(\lambda, \mu, \sigma) \Big) \end{aligned}$$

where $I(\lambda, \mu, \sigma)$ is Fisher-information matrix and it is given by

$$I(\lambda, \mu, \sigma) = \begin{pmatrix} I(\lambda) & I(\lambda, \mu) & I(\lambda, \sigma) \\ I(\lambda, \mu) & I(\mu) & I(\mu, \sigma) \\ I(\lambda, \sigma) & I(\mu, \sigma) & I(\sigma) \end{pmatrix},$$

such that

$$I(\lambda) = 7\sigma^2(x)/4$$

$$
\begin{aligned}
I(\mu) &= (1+2\lambda(x)^2\sigma(x)^2)/(\mu(x)^2\sigma(x)^2) \\
I(\sigma) &= 2/\sigma(x)^2 \\
I(\lambda,\mu) &= -\frac{1}{2\mu(x)} \\
I(\lambda,\sigma) &= \lambda(x)\sigma(x) \\
I(\mu,\sigma) &= \frac{2\lambda(x)}{\mu(x)\sigma(x)}
\end{aligned}
$$

$$
\begin{aligned}
a(x) &= \sigma(x)m''(x;\lambda,\mu,\sigma) \\
b(x) &= \frac{m''(x;\lambda,\mu,\sigma)}{\mu(x)\sigma(x)}
\end{aligned}
$$

with

$$
m(x;\lambda,\mu,\sigma) \approx (\lambda(x)-1)\sigma(x) \tag{5.15}
$$

Proof: We still use Corollary 5.1. To calculate Fisher-Information matrix, let $\theta(x) = (\lambda(x), \mu(x), \sigma(x))^T$ be the parametric function of $\theta = (\lambda, \mu, \sigma)^T$, and L is the log density function of Y: $L = \lambda log\frac{Y}{\mu} - log\sigma - 1/2Z(c)^2$, then

$$
I(\theta(x)) = -E\frac{\partial^2 L}{\partial\theta^2}|_{\theta=\theta(x)} =
$$

$$
\begin{pmatrix}
E(\frac{\partial L}{\partial\lambda}|_{\theta=\theta(x)})^2 & -E\frac{\partial^2 L}{\partial\lambda\mu}|_{\theta=\theta(x)} & -E\frac{\partial^2 L}{\partial\lambda\sigma}|_{\theta=\theta(x)} \\
-E\frac{\partial^2 L}{\partial\lambda\mu}|_{\theta=\theta(x)} & E(\frac{\partial L}{\partial\mu}|_{\theta=\theta(x)})^2 & -E\frac{\partial^2 L}{\partial\mu\sigma}|_{\theta=\theta(x)} \\
-E\frac{\partial^2 L}{\partial\lambda\sigma}|_{\theta=\theta(x)} & -E\frac{\partial^2 L}{\partial\mu\sigma}|_{\theta=\theta(x)} & E(\frac{\partial L}{\partial\sigma}|_{\theta=\theta(x)})^2
\end{pmatrix}.
$$

Since

$$
\begin{aligned}
\frac{\partial L}{\partial\sigma} &= \frac{(Z^2-1)}{\sigma} \\
\frac{\partial L}{\partial\mu} &= \frac{Z}{\mu\sigma} + \frac{\lambda(Z^2-1)}{\mu} \\
\frac{\partial L}{\partial\lambda} &= \frac{Z}{\lambda}(Z - \frac{log(Y/\mu)}{\sigma}) - log(Y/\mu)(Z^2-1)
\end{aligned}
$$

and as $E\frac{\partial L}{\partial\lambda}$ is very complicated this may be approximated by a Taylor expansion of $log\frac{Y}{\mu}$ i.e either the transformation power λ or the coefficient of variation σ is

small which indeed matches most of practical situations we experimented and knew (e.g. figures in Chapter 5 and figures in Cole (1988) and Cole and Green (1992)):

$$log\frac{Y}{\mu} = \frac{1}{\lambda}log(\frac{Y}{\mu})^{\lambda} = \frac{1}{\lambda}log\{1+\lambda\sigma Z\} \approx \sigma Z - \frac{1}{2}\lambda\sigma^2 Z^2,$$

and Z is N(0, 1) variable, for example,

$$\begin{aligned}\frac{\partial L}{\partial \lambda} &\approx \sigma Z - \frac{1}{2}\sigma^2\lambda Z^2 \\ \frac{\partial^2 L}{\partial \sigma^2} &= -\frac{Z^2+1}{\sigma^2}\end{aligned}$$

Let

$$\begin{aligned}l_\lambda(x) = \frac{\partial L_c}{\partial \lambda} &= \sum_i \Big(\frac{Z_i}{\lambda}(Z_i - \frac{log(Y_i/\mu)}{\sigma}) \\ &- log(Y_i/\mu)(Z_i^2-1)\Big)K(\frac{X_i - x}{h}), \\ l_\mu(x) = \frac{\partial L_c}{\partial \mu} &= \sum_i \Big(\frac{Z_i}{\mu\sigma} + \frac{\lambda(Z_i^2-1)}{\mu}\Big)K(\frac{X_i - x}{h}), \\ l_\sigma(x) = \frac{\partial L_c}{\partial \sigma} &= \sum_i \Big(\frac{(Z_i^2-1)}{\sigma}\Big)K(\frac{X_i - x}{h}).\end{aligned}$$

From

$$\begin{aligned}\frac{\partial Z}{\partial \lambda} &= \frac{1}{\lambda}(\frac{log(Y/\mu)}{\sigma} - Z) + log(Y/\mu)Z \\ \frac{\partial Z}{\partial \mu} &= -\frac{\lambda}{\mu}Z - \frac{1}{\mu\sigma} \\ \frac{\partial Z}{\partial \sigma} &= -\frac{Z}{\sigma}\end{aligned}$$

and Taylor expansion of $m_2(X;\lambda,\mu,\sigma) = E\{Z^2|X\}$, and noting that $m_2(x;\lambda,\mu,\sigma) = 1$, we have

$$\begin{aligned}a_n^2 El_\sigma(x) &= \sum_i \Big(\frac{m_2'(x;\lambda,\mu,\sigma)(X_i-x) + \frac{1}{2}m_2''(x;\lambda,\mu,\sigma)(X_i-x)^2}{\sigma(x)}\Big)K(\frac{X_i-x}{h}) \\ &+ o(h^2) \\ &= \frac{1}{2\sigma(x)}m_2''(x;\lambda,\mu,\sigma)g(x)h^2\mu_2(K) + o(h^2).\end{aligned}$$

Similarly, Taylor expansion of $m(X;\lambda,\mu,\sigma) = E\{Z|X\}$, and noting that $m(x;\lambda,\mu,\sigma) = 0$,

$$a_n^2 El_\mu(x) = \frac{1}{2}\Big(\frac{m''(x;\lambda,\mu,\sigma)}{\mu(x)\sigma(x)} + \frac{\lambda(x)}{\mu(x)} m_2''(x;\lambda,\mu,\sigma)\Big) h^2 \mu_2(K) g(x) + o(h^2),$$

$$a_n^2 El_\lambda(x) = \frac{1}{2}\Big(\sigma(x) m''(x;\lambda,\mu,\sigma) - \frac{1}{2}\lambda(x)\sigma^2(x) m_2''(x;\lambda,\mu,\sigma)\Big) h^2 \mu_2(K) g(x) + o(h^2).$$

Now we are in a position of trying to simplify $m''(x;\lambda,\mu,\sigma)$ and $m_2''(x;\lambda,\mu,\sigma)$, starting from

$$m(X;\lambda,\mu,\sigma) = E\{Z(c)\} = E\frac{(Y/\mu(X))^{\lambda(X)} - 1}{\lambda(X)\sigma(X)},$$

and

$$m_2(X;\lambda,\mu,\sigma) = E\{Z(c)^2\} = E\Big(\frac{(Y/\mu(X))^{\lambda(X)} - 1}{\lambda(X)\sigma(X)}\Big)^2.$$

Firstly regard $\frac{(Y/\mu(X))^{\lambda(X)}-1}{\lambda(X)\sigma(X)}$ as $Y's$ function, $f(Y)$, for each X; then expanding $f(Y)$ at $E(Y) = \mu(X)$:

$$f(Y) = f(\mu(X)) + f'(\mu)(Y - \mu(X)) + \frac{1}{2} f''(\mu(X))(Y - \mu(X))^2 + ...,$$

and noting that only $Y's$ second moment is involved in LMS method, so

$$Ef(Y) \approx f(\mu(X)) + \frac{1}{2} f''(\mu(X))(Y - \mu(X))^2.$$

From

$$f''(Y) = \frac{(\lambda(X) - 1)(Y/\mu(X))^{\lambda(X)-2}\frac{1}{\mu(X)^2}}{\sigma(X)},$$

and according to the definition of LMS method,

$$E\Big(\frac{Y}{\mu(X)} - 1\Big)^2 = \sigma(X)^2.$$

Thus

$$m(X;\lambda,\mu,\sigma) \approx \frac{1}{2}(\lambda(X) - 1)\sigma(X)$$

and from $m''(x;\lambda,\mu,\sigma) = m''(X;\lambda,\mu,\sigma)|_{X=x}$, we have

$$m''(x;\lambda,\mu,\sigma) \approx \frac{1}{2}\lambda''(x)\sigma(x) + \frac{1}{2}\sigma''(x)(\lambda(x)-1) + \lambda'(x)\sigma'(x).$$

Similarly, let

$$f(Y) = \frac{(Y/\mu(X))^{2\lambda(X)} - 2(Y/\mu(X))^{\lambda(X)} + 1}{\lambda(X)^2\sigma(X)^2},$$

and

$$f''(Y) = \frac{2(2\lambda(X)-1)(Y/\mu(X))^{2\lambda(X)-2}\frac{1}{\mu(X)^2} - 2(\lambda(X)-1)(Y/\mu(X))^{\lambda(X)-2}\frac{1}{\mu(X)^2}}{\lambda(X)\sigma(X)^2},$$

we have

$$m_2(X;\lambda,\mu,\sigma) \approx 1$$

and

$$m_2''(x;\lambda,\mu,\sigma) \approx 0.$$

Further, if let $Var(u;\lambda,\mu,\sigma) = Var\{Z^2|X=u\}$, then under the condition ("true model" is normal once λ, μ and σ transformation are done) of theorem, we have:

$$a_n^2 Var\{l_\sigma(x)\} = a_n^2 \sum_i \frac{Var\{Z_i^2\}}{\sigma(x)^2} K^2(\frac{X_i - x}{h})$$

$$= h^{-1}\frac{1}{\sigma(x)^2}\int var(u;\lambda,\mu,\sigma)K^2(\frac{u-x}{h})g(u)du$$

$$= \frac{1}{\sigma(x)^2}\int var(x+uh;\lambda,\mu,\sigma)K^2(u)g(x+uh)du$$

$$= \frac{1}{\sigma(x)^2}\int var(x;\lambda,\mu,\sigma)K^2(u)g(x)du + o(h)$$

$$= \frac{2}{\sigma(x)^2}g(x)R(K) + o(h) = R(K)g(x)I(\sigma) + o(h)$$

Similar expression are obtained for other components such as $a_n^2 Var\{l_\sigma(x)\}$ and $a_n^2 Var\{l_\lambda(x)\}$ and covariances of $a_n^2 Var\{\frac{\partial L_c}{\partial\theta(x)}\}$, and eventually,

$$a_n^2 Var\{\frac{\partial L_c}{\partial\theta(x)}\} \approx R(K)g(x)I(\lambda(x),\mu(x),\sigma(x)).$$

5.4.2 Practical Computation

The estimates of λ, μ and σ calculated by first solving $l_\sigma(x) = 0$ for σ^2 in terms of μ and λ to obtain

$$\hat{\sigma}^2 = \frac{\sum((y_i/\mu)^\lambda - 1)^2 K(\frac{x-x_i}{h})}{\lambda^2 \sum K(\frac{x-x_i}{h})} \tag{5.16}$$

Then substitute in $l_\mu(x) = 0$ and solve for μ^λ to get

$$\mu^\lambda = \frac{\sum(y_i)^\lambda K(\frac{x-x_i}{h})}{\sum K(\frac{x-x_i}{h})} \tag{5.17}$$

and finally substitute in $l_\lambda(x) = 0$ to obtain the following relation for λ,

$$\lambda = \frac{\sum((y_i/\mu)^l - 1)^2 K(\frac{x-x_i}{h})}{\sum\Big(((y_i/\mu)^l - 1)^2 + (y_i/\mu)^l - 1\Big) log(y_i/\mu) K(\frac{x-x_i}{h}) - l^2 s^2 \sum log(y_i/\mu) K(\frac{x-x_i}{h})} \tag{5.18}$$

Under this algorithm, first step of iteration is carried out starting from initial value for λ, say $\lambda = \lambda_0 = 1$, then substitute in (5.17) to obtain a solution for $\mu = \mu_0$, and in turn substitute in (5.16) and solve for σ^2 (say $\sigma^2 = \sigma_0^2$), finally insert λ_0, μ_0 and σ_0^2 in the equation (5.18) and result in the $1st$ value of $\lambda = \lambda_1$. Eventually, the solutions for λ, μ and σ are obtained iteratively by cycling around the three equations.

Advantages of this computational process are no matrix computations are involved and it converges faster than Fisher-Scoring algorithm particularly for M and S curves.

For example, for all sets of data used in this thesis, three iterations give a stable solution for M and S, but not so stable for L, however, for estimating quantile curve using L, M, and S just need three or a bit more iterations.

5.4.3 Local Linear Fitting LMS

Local linear kernel fitting for L, M and S curves is investigated. On writing the kernel weighted log-likelihood function L_l in terms of $\lambda+(x_i-x)\lambda_1$, $\mu+(x_i-x)\mu_1$ and $\sigma+(x_i-x)\sigma_1$ instead of λ, μ and σ, the variable Z takes the form now

$$Z(l) = \frac{[Y_i/(\mu+(x_i-x)\mu_1)]^{\lambda+(x_i-x)\lambda_1}-1}{(\lambda+(x_i-x)\lambda_1)(\sigma+(x_i-x)\sigma_1)} \tag{5.19}$$

and log-likelihood function is

$$\begin{aligned} L(l) &= \sum_i \Big((\lambda+(x_i-x)\lambda_1) log \frac{Y_i}{\mu+(x_i-x)\mu_1} \\ &- log(\sigma+(x_i-x)\sigma_1) - 1/2 Z(l)_i^2 \Big) K((x-x_i)/h) \end{aligned} \tag{5.20}$$

Theorem 5.5: Under the regular conditions, with local linear fitting,

$$\begin{aligned} \sqrt{nh}\Big[& \Big(L(x)-\lambda(x), M(x)-\mu(x), S(x)-\sigma(x) \Big)^T \\ & - \frac{1}{2}h^2\mu_2(K) I^{-1}(\lambda,\mu,\sigma)(\lambda(x)-1)\sigma''(x)+\lambda''(x)\sigma(x)) \Big(a(x), b(x), 0 \Big)^T \Big] \\ & \rightarrow N\Big(0, \frac{R(K)}{g(x)} I^{-1}(\lambda,\mu,\sigma)\Big) \end{aligned}$$

where $I(\lambda,\mu,\sigma)$ is Fisher-information matrix and it is given by

$$I(\lambda,\mu,\sigma) = \begin{pmatrix} I(\lambda) & I(\lambda,\mu) & I(\lambda,\sigma) \\ I(\lambda,\mu) & I(\mu) & I(\mu,\sigma) \\ I(\lambda,\sigma) & I(\mu,\sigma) & I(\sigma) \end{pmatrix},$$

as at Theorem 5.4 and

$$a(x) = \sigma(x)$$

$$b(x) = \frac{1}{\mu(x)\sigma(x)}$$

Remark. Based on same approximate order ($o(h^2)$), the main difference Theorem 5.4 and Theorem 5.5 lines in the bias term of local linear fitting does not depend on parameters' first derivative function.

Proof: Following same lines as the proof for local constant fitting of Theorem 5.4 , but now

$$\begin{aligned} l_\lambda(x) = \frac{\partial L(l)}{\partial \lambda} &= \sum_i \Big(\frac{Z(l)_i}{\lambda + (x_i - x)\lambda_1} (Z(l)_i - \frac{log(Y_i/(\mu + (x_i - x)\mu_1))}{\sigma + (x_i - x)\sigma_1}) \\ &- log(Y_i/\mu + (x_i - x)\mu_1)(Z(l)_i^2 - 1) \Big) K(\frac{X_i - x}{h}) \\ &\approx \sum_i (\sigma + (x_i - x)\sigma_1) Z(l)_i K(\frac{X_i - x}{h}) \\ &- \frac{1}{2} \sum_i (\lambda + (x_i - x)\lambda_1)(\sigma + (x_i - x)\sigma_1)^2 Z(l)_i^2 K(\frac{X_i - x}{h}) \\ l_\mu(x) = \frac{\partial L(l)}{\partial \mu} &= \sum_i \Big(\frac{Z(l)_i}{(\mu + (x_i - x)\mu_1)(\sigma + (x_i - x)\sigma_1)} \\ &+ \frac{(\lambda + (x_i - x)\lambda_1)(Z(l)_i^2 - 1)}{\mu + (x_i - x)\mu_1} \Big) K(\frac{X_i - x}{h}) \\ l_\sigma(x) = \frac{\partial L(l)}{\partial \sigma} &= \sum_i \Big(\frac{(Z(l)_i^2 - 1)}{\sigma + (x_i - x)\sigma_1} \Big) K(\frac{X_i - x}{h}) \end{aligned}$$

so

$$a_n^2 E l_\sigma(x) =$$

$$\sum_i \Big(\frac{m_2'(x_i - x; \lambda, \mu, \sigma)(X_i - x) + \frac{1}{2} m_2''(x_i - x; \lambda, \mu, \sigma)(X_i - x)^2}{\sigma(x)} \Big) K(\frac{X_i - x}{h}) + o(h^2)$$

where

$$m_2(x_i - x; \lambda, \mu, \sigma) = E\{Z(l)_i^2 | X = x\} \approx 1. \tag{5.21}$$

Thus

$$a_n^2 E l_\sigma(x) = 0 + o(h^2).$$

Similarly, from

$$a_n^2 E l_\mu(x) = \sum_i \frac{(x_i - x) m'(x_i - x; \lambda, \mu, \sigma) + \frac{1}{2}(x_i - x)^2 m''(x_i - x; \lambda, \mu, \sigma)}{\mu(x)\sigma(x)} K(\frac{X_i - x}{h})$$

$$\begin{aligned} a_n^2 E l_\lambda(x) &= a_n^2 \sum_i (\sigma + (x_i - x)\sigma_1)((x_i - x) m'(x_i - x; \lambda, \mu, \sigma) \\ &+ \frac{1}{2}(x_i - x)^2 m''(x_i - x; \lambda, \mu, \sigma)) K(\frac{X_i - x}{h}) \end{aligned}$$

where

$$m(x_i - x; \lambda, \mu, \sigma) = E\{Z(l)_i | X = x\}$$

$$
\begin{aligned}
& \approx (\lambda(x) - 1 + (x_i - x)\lambda_1(x))(\sigma(x) + (x_i - x)\sigma_1(x)) \\
m'(x_i - x; \lambda, \mu, \sigma) & \approx (x_i - x)\sigma''(x)\Big(\lambda(x) - 1 + (x_i - x)\lambda_1(x)\Big) \\
& + (x_i - x)\lambda''(x)\Big(\sigma(x) - 1 + (x_i - x)\sigma_1(x)\Big) \\
& = (\lambda(x) - 1)\sigma''(x)(x_i - x) + (x_i - x)^2\lambda_1(x)\sigma''(x) \\
& + (x_i - x)\lambda''(x))\sigma(x) + (x_i - x)^2\lambda''(x)\sigma_1'(x))
\end{aligned}
$$

$$
\begin{aligned}
m''(x_i - x; \lambda, \mu, \sigma) &= \lambda'(x)\sigma''(x) + (\lambda(x) - 1)\sigma'''(x)(x_i - x) \\
&- (\lambda(x) - 1)\sigma''(x) - 2\lambda_1(x)\sigma''(x)(x_i - x) \\
&+ (x_i - x)^2\lambda_1'(x)\sigma''(x) + (x_i - x)^2\lambda_1(x)\sigma'''(x) \\
&- \lambda''(x)\sigma(x) + (x_i - x)\lambda'''(x)\sigma(x) + (x_i - x)\lambda''(x)\sigma'(x) \\
&- 2(x_i - x)\lambda''(x)\sigma_1(x) + (x_i - x)^2\lambda'''(x)\sigma_1(x) + (x_i - x)^2\lambda''(x)\sigma_1'(x)
\end{aligned}
$$

Thus,

$$
a_n^2 El_\mu(x) = \frac{(\lambda(x) - 1)\sigma''(x) + \lambda''(x)\sigma(x)}{2}\Big(\frac{1}{\mu(x)\sigma(x)}\Big)h^2\mu_2(K)g(x) + o(h^2),
$$

$$
a_n^2 El_\lambda(x) = \frac{1}{2}\sigma(x)\Big((\lambda(x) - 1)\sigma''(x) + \lambda''(x)\sigma(x)\Big)h^2\mu_2(K)g(x) + o(h^2).
$$

Here, we use through, for bounded functions a, b and an integer c,

$$
\int \frac{t^c K(t)}{a + bht}dt = \frac{1}{a}\int t^c K(t)dt + O(h).
$$

and we have

$$
a_n^2 Var\{\frac{\partial L_c}{\partial\theta(x)}\} \approx R(K)g(x)I(\lambda(x), \mu(x), \sigma(x)).
$$

Hence Theorem 5.5.

Now, according from section 5.2.3, and note that

$$
q_p(x) = Q_{\theta(x)}^{-1}(\Phi^{-1}(p)) = \mu(x)(\lambda(x)\sigma(x)\Phi^{-1}(p) + 1)^{1/\lambda(x)}
$$

and

$$
\frac{\partial Q_{\theta(x)}^{-1}(\Phi^{-1}(p))}{\partial\theta(x)} = \frac{\partial q_p(x)}{\partial\theta(x)} = \Big(\frac{\partial q_p(x)}{\partial\lambda(x)}, \frac{\partial q_p(x)}{\partial\mu(x)}, \frac{\partial q_p(x)}{\partial\sigma(x)}\Big)^T
$$

such that

$$\begin{aligned}
\frac{\partial Q^{-1}_{\theta(x)}(\Phi^{-1}(p))}{\partial \lambda(x)} &= q_p(x)\Big(\frac{log\frac{\mu(x)}{q_p(x)}}{\lambda(x)} + \frac{\sigma(x)\Phi^{-1}(p)}{\lambda(x)(1+\lambda(x)\sigma(x)\Phi^{-1}(p))}\Big) \\
\frac{\partial Q^{-1}_{\theta(x)}(\Phi^{-1}(p))}{\partial \mu(x)} &= \frac{q_p(x)}{\mu(x)} \\
\frac{\partial Q^{-1}_{\theta(x)}(\Phi^{-1}(p))}{\partial \sigma(x)} &= q_p(x)\frac{\Phi^{-1}(p)}{1+\lambda(x)\sigma(x)\Phi^{-1}(p)}
\end{aligned}$$

Theorem 5.6 For local constant fitting or local linear fitting and under regularity conditions

$$\begin{aligned}
&\sqrt{nh}\ \Big[\Big(\hat{q}_p(x) - q_p(x)\Big)^T \\
&- \ h^2\mu_2(K)\Big(\frac{\partial Q^{-1}_{\theta(x)}(\Phi^{-1}(p))}{\partial \lambda(x)}\Big)^T\Big(a(x), b(x), 0\Big)\Big] \\
&\to \ N\Big(0, \frac{R(K)}{g(x)}\Big(\frac{\partial Q^{-1}_{\theta(x)}(\Phi^{-1}(p))}{\partial \lambda(x)}\Big)^T I^{-1}(\lambda,\mu,\sigma)\Big(\frac{\partial Q^{-1}_{\theta(x)}(\Phi^{-1}(p))}{\partial \lambda(x)}\Big)^T\Big)
\end{aligned}$$

and asymptotic mean square error

$$\begin{aligned}
MSE(\hat{q}_p(x)) = &\Big[h^2\mu_2(K)\Big(\frac{\partial Q^{-1}_{\theta(x)}(\Phi^{-1}(p))}{\partial \lambda(x)}\Big)^T\Big(a(x), b(x), 0\Big)\Big]^2 \\
+ \quad &\frac{R(K)}{g(x)nh}\Big(\frac{\partial Q^{-1}_{\theta(x)}(\Phi^{-1}(p))}{\partial \lambda(x)}\Big)^T I^{-1}(\lambda,\mu,\sigma)\Big(\frac{\partial Q^{-1}_{\theta(x)}(\Phi^{-1}(p))}{\partial \lambda(x)}\Big)^T.
\end{aligned}$$

5.4.4 The Practical Computation of Local Linear Fitting

In this section, a computation technique is discussed, first using local constant fitting to obtain an initial estimator which in turn is plugged in the bandwidth selection method for local linear approach.

However, as it is required that the response $\frac{y_i}{\mu+(t_i-t)\mu_1}$ is positive for any observation (y_i, t_i), and most of age-related data sets have response Y positive, so without loss of generality, the positivity of $\mu + (t_i - t)\mu_1$ is required in our algorithm after a local linear fitting.

Thus, given t based on the observations $\{t_i\}_1^n$, define

$$z_i^* = \frac{[\frac{y_i}{((\mu+(t_i-t)\mu_1)I(\mu+(t_i-t)\mu_1>0)}]^{\lambda+(t_i-t)\lambda_1} - 1}{(\lambda+(t_i-t)\lambda_1)(\sigma+(t_i-t)\sigma_1)} \tag{5.22}$$

This corresponds to the log-likelihood function

$$\begin{aligned} L^* &= \sum_i \Big((\lambda+(t_i-t)\lambda_1) log \frac{y_i}{(\mu+(t_i-t)\mu_1)I(\mu+(t_i-t)\mu_1>0)} \\ &- log(\sigma+(t_i-t)\sigma_1) - 1/2 z_i^{*2} \Big) K((t-t_i)/h) \end{aligned} \tag{5.23}$$

Maximizing L^* results in six estimators $(l,l_1),(m,m_1),(s,s_1)$ for parameters $(\lambda,\lambda_1);(\mu,\mu_1);(\sigma,\sigma_1)$, and the l, m and s are just $L(t), M(t)$ and $S(t)$ while l_1, m_1, and s_1 estimate respectively the derivatives of $L(t)$, $M(t)$ and $S(t)$.

Unlike computing the local constant fitting, this estimation falls within the compas of the Fisher-Scoring computing rule.

Let $M_i = (\mu+(t_i-t)\mu_1)I(\mu+(t_i-t)\mu_1>0)$.

Note that

$$\begin{aligned} \frac{\partial z_i^*}{\partial \lambda} &= \frac{1}{\lambda+(t_i-t)\lambda_1}(\frac{log(y_i/M_i)}{\sigma+(t_i-t)\sigma_1} - z_i^*) \\ &+ log(y_i/M_i) z_i^* \end{aligned} \tag{5.24}$$

$$\frac{\partial z_i^*}{\partial \lambda_1} = (t_i-t)\frac{\partial z_i^*}{\partial \lambda} \tag{5.25}$$

$$\frac{\partial z_i^*}{\partial \mu} = -\frac{\lambda+(t_i-t)\lambda_1}{M_i} z_i^* - \frac{1}{M_i(\sigma+(t_i-t)\sigma_1)} \tag{5.26}$$

$$\frac{\partial z_i^*}{\partial \mu_1} = (t_i-t)\frac{\partial z_i^*}{\partial \mu} \tag{5.27}$$

$$\frac{\partial z_i^*}{\partial \sigma} = -\frac{z_i^*}{\sigma+(t_i-t)\sigma_1} \tag{5.28}$$

$$\frac{\partial z_i^*}{\partial \sigma_1} = (t_i-t)\frac{\partial z_i^*}{\partial \sigma} \tag{5.29}$$

If expressing simply $\sum_i (t_i-t).U_i$ as $(t_i-t)\sum_i .U_i$, then

$$l_\lambda(t) = \frac{\partial L^*}{\partial \lambda} = \sum_i \Big(\frac{z_i^*}{\lambda+(t_i-t)\lambda_1}(z_i^* - \frac{log(y_i/M_i)}{\sigma+(t_i-t)\sigma_1}) \Big)$$

$$- \quad log(y_i/M_i)(z_i^{*2}-1)\Big)K(\frac{t-t_i}{h}) \tag{5.30}$$

$$l_{\lambda_1}(t) = \frac{\partial L^*}{\partial \lambda_1} \quad = \quad (t_i - t)l_\lambda(t) \tag{5.31}$$

$$l_\mu(t) = \frac{\partial L^*}{\partial \mu} \quad = \quad \sum_i \Big(\frac{z_i^*}{M_i(\sigma + (t_i - t)\sigma_1)}$$

$$+ \quad \frac{(\lambda + (t_i - t)\lambda_1)(z_i^{*2}-1)}{M_i}\Big)K(\frac{t-t_i}{h}) \tag{5.32}$$

$$l_{\mu_1}(t) = \frac{\partial L^*}{\partial \mu_1} \quad = \quad (t_i - t)l_\mu(t) \tag{5.33}$$

$$l_\sigma(t) = \frac{\partial L^*}{\partial \sigma} \quad = \quad \sum_i \Big(\frac{(z_i^{*2}-1)}{\sigma + (t_i - t)\sigma_1}\Big)K(\frac{t-t_i}{h}) \tag{5.34}$$

$$l_{\sigma_1}(t) = \frac{\partial L^*}{\partial \sigma_1} \quad = \quad (t_i - t)l_\sigma(t) \tag{5.35}$$

The matrix $E(-\frac{\partial^2 L^*}{\partial \beta \partial \beta^T}) = W$ is a 6×6 symmetric matrix with rows

$$c_1 = (W_\lambda, W_{\lambda\lambda_1}, W_{\lambda\mu}, W_{\lambda\mu_1}, W_{\lambda\sigma}, W_{\lambda\sigma_1})$$

$$c_2 = (W_{\lambda\lambda_1}, W_{\lambda_1}, W_{\lambda_1\mu}, W_{\lambda_1\mu_1}, W_{\lambda_1\sigma}, W_{\lambda_1\sigma_1})$$

$$c_3 = (W_{\lambda\mu}, W_{\lambda_1\mu}, W_\mu, W_{\mu\mu_1}, W_{\mu\sigma}, W_{\mu\sigma_1})$$

$$c_4 = (W_{\lambda\mu_1}, W_{\lambda_1\mu_1}, W_{\mu\mu_1}, W_{\mu_1}, W_{\mu_1\sigma}, W_{\mu_1\sigma_1})$$

$$c_5 = (W_{\lambda\sigma}, W_{\lambda_1\sigma}, W_{\mu\sigma}, W_{\mu_1\sigma}, W_\sigma, W_{\sigma\sigma_1})$$

$$c_6 = (W_{\lambda\sigma_1}, W_{\lambda_1\sigma_1}, W_{\mu\sigma_1}, W_{\mu_1\sigma_1}, W_{\sigma\sigma_1}, W_{\sigma_1})$$

where

$$W_\lambda = -E\{\frac{\partial^2 L^*}{\partial \lambda^2}\} = \frac{7}{4}\sum(\sigma + (t_i - t)\sigma_1)^2 K(\frac{t_i - t}{h})$$

$$W_{\lambda\lambda_1} = -E\{\frac{\partial^2 L^*}{\partial \lambda \partial \lambda_1}\} = \frac{7}{4}\sum(t_i - t)(\sigma + (t_i - t)\sigma_1)^2 K(\frac{t_i - t}{h})$$

$$W_{\lambda\mu} = -E\{\frac{\partial^2 L^*}{\partial \lambda \partial \mu}\} = -\sum(1/M_i)K(\frac{t_i - t}{h})$$

$$W_{\lambda\mu_1} = -E\{\frac{\partial^2 L^*}{\partial \lambda \partial \mu_1}\} = -\sum((t_i - t)/M_i)K(\frac{t_i - t}{h})$$

$$W_{\lambda\sigma} = -E\{\frac{\partial^2 L^*}{\partial \lambda \partial \sigma}\} = \sum(\lambda + (t_i - t)\lambda_1)(\sigma + (t_i - t)\sigma_1)K(\frac{t_i - t}{h})$$

$$W_{\lambda\sigma_1} = -E\{\frac{\partial^2 L^*}{\partial \lambda \partial \sigma_1}\} = \sum(t_i - t)(\lambda + (t_i - t)\lambda_1)(\sigma + (t_i - t)\sigma_1)K(\frac{t_i - t}{h})$$

$$
\begin{aligned}
W_{\lambda_1} &= -E\{\frac{\partial^2 L^*}{\partial \lambda_1^2}\} = \frac{7}{4}\sum (t_i - t)^2(\sigma + (t_i - t)\sigma_1)^2 K(\frac{t_i - t}{h}) \\
W_{\lambda_1 \mu} &= -E\{\frac{\partial^2 L^*}{\partial \lambda_1 \partial \mu}\} = W_{\lambda \mu_1} \\
W_{\lambda_1 \mu_1} &= -E\{\frac{\partial^2 L^*}{\partial \lambda_1 \partial \mu_1}\} = -\sum ((t_i - t)^2 / M_i) K(\frac{t_i - t}{h}) \\
W_{\lambda_1 \sigma} &= -E^*\{\frac{\partial^2 L^*}{\partial \lambda_1 \partial \sigma}\} = W_{\lambda \sigma_1} \\
W_{\lambda_1 \sigma_1} &= -E\{\frac{\partial^2 L^*}{\partial \lambda_1 \partial \sigma_1}\} = \sum (t_i - t)^2(\lambda + (t_i - t)\lambda_1)(\sigma + (t_i - t)\sigma_1) K(\frac{t_i - t}{h}) \\
W_{\mu} &= -E\{\frac{\partial^2 L^*}{\partial \mu^2}\} = \sum \frac{\left(1 + 2(\lambda + (t_i - t)\lambda_1)^2(\sigma + (t_i - t)\sigma_1)^2\right)}{\left(M_i^2(\sigma + (t_i - t)\sigma_1)^2\right)} K(\frac{t_i - t}{h}) \\
W_{\mu \mu_1} &= -E\{\frac{\partial^2 L^*}{\partial \mu \partial \mu_1}\} \\
&= \sum \frac{\left(1 + 2(\lambda + (t_i - t)\lambda_1)^2(\sigma + (t_i - t)\sigma_1)^2\right)(t_i - t)}{\left(M_i^2(\sigma + (t_i - t)\sigma_1)^2\right)} K(\frac{t_i - t}{h}) \\
W_{\mu \sigma} &= -E\{\frac{\partial^2 L^*}{\partial \mu \partial \sigma}\} = 2\sum \frac{(\lambda + (t_i - t)\lambda_1)}{M_i(\sigma + (t_i - t)\sigma_1)} K(\frac{t_i - t}{h}) \\
W_{\mu \sigma_1} &= -E\{\frac{\partial^2 L^*}{\partial \mu \partial \sigma_1}\} = 2\sum \frac{(\lambda + (t_i - t)\lambda_1)(t_i - t)}{M_i(\sigma + (t_i - t)\sigma_1)} K(\frac{t_i - t}{h}) \\
W_{\mu_1} &= -E\{\frac{\partial^2 L^*}{\partial \mu_1^2}\} = 2\sum \frac{(\lambda + (t_i - t)\lambda_1)(t_i - t)^2}{M_i(\sigma + (t_i - t)\sigma_1)} K(\frac{t_i - t}{h}) \\
W_{\mu_1 \sigma} &= -E\{\frac{\partial^2 L^*}{\partial \mu_1 \partial \sigma}\} = W_{\mu \sigma_1} \\
W_{\mu_1 \sigma_1} &= -E\{\frac{\partial^2 L^*}{\partial \mu_1 \partial \sigma_1}\} = 2\sum \frac{(\lambda + (t_i - t)\lambda_1)(t_i - t)^2}{M_i(\sigma + (t_i - t)\sigma_1)} K(\frac{t_i - t}{h}) \\
W_{\sigma} &= -E\{\frac{\partial^2 L^*}{\partial \sigma^2}\} = 2\sum \frac{1}{(\sigma + (t_i - t)\sigma_1)^2} K(\frac{t_i - t}{h}) \\
W_{\sigma \sigma_1} &= -E\{\frac{\partial^2 L^*}{\partial \sigma^2 \partial \sigma_1}\} = 2\sum \frac{(t_i - t)}{(\sigma + (t_i - t)\sigma_1)^2} K(\frac{t_i - t}{h}) \\
W_{\sigma_1} &= -E\{\frac{\partial^2 L^*}{\partial \sigma^2}\} = 2\sum \frac{(t_i - t)^2}{(\sigma + (t_i - t)\sigma_1)^2} K(\frac{t_i - t}{h})
\end{aligned}
$$

Note that there are only 36-15-3=18 independent elements in this matrix. Also, the small λ or σ approximation has been used.

As an example of calculations,

$$\begin{aligned} W_\sigma &= -E\frac{\partial L^*}{\partial \sigma} = E^* \sum_i \frac{z_i^{*2}+1}{(\sigma+(t_i-t)\sigma_1)^2} K(\frac{t_i-t}{h}) \\ &\approx 2\sum \frac{1}{(\sigma+(t_i-t)\sigma_1)^2} K(\frac{t_i-t}{h}) \end{aligned} \tag{5.36}$$

Only 6 functions are needed to run the 36 elements of matrix W when S-plus is used, i.e $\left(W_\lambda, W_{\lambda\lambda 1}, W_{\lambda 1}\right)$ can be solved using one function with $k = 0, 1, 2$ which corresponds to the power of $(t_i - t)^k$. Similarly, one function for each of $\left(W_{\lambda\mu}, W_{\lambda\mu 1}, W_{\lambda 1\mu 1}\right)$, $\left(W_{\lambda\sigma}, W_{\lambda\sigma 1}, W_{\lambda 1\sigma 1}\right)$, $\left(W_\mu, W_{\mu\mu 1}, W_{\mu 1}\right)$, $\left(W_{\mu\sigma}, W_{\mu\sigma 1}, W_{\mu 1\sigma 1}\right)$ and $\left(W_\sigma, W_{\sigma\sigma 1}, W_{\sigma 1}\right)$.

The six components of the vector $\frac{\partial l}{\partial \beta}$ are

$$\begin{aligned} \frac{\partial l}{\partial \lambda} &= \sum_i \frac{z_i}{\lambda+(t_i-t)\lambda_1}\left(z_i - \frac{log(\frac{y_i}{M_i})}{\sigma+(t_i-t)\sigma_1} - log(\frac{y_i}{M_i})\right) K(\frac{t_i-t}{h}) \\ \frac{\partial l}{\partial \lambda_1} &= \sum_i \frac{z_i(t_i-t)}{\lambda+(t_i-t)\lambda_1}\left(z_i - \frac{log(\frac{y_i}{M_i})}{\sigma+(t_i-t)\sigma_1} - log(\frac{y_i}{M_i})\right) K(\frac{t_i-t}{h}) \\ \frac{\partial l}{\partial \mu} &= \sum_i \left(\frac{z_i}{M_i(\sigma+(t_i-t)\sigma_1)} + \frac{(\lambda+(t_i-t)\lambda_1)(z_i^2-1)}{M_i}\right) K(\frac{t_i-t}{h}) \\ \frac{\partial l}{\partial \mu_1} &= \sum_i (t_i-t)\left(\frac{z_i}{M_i(\sigma+(t_i-t)\sigma_1)} + \frac{(\lambda+(t_i-t)\lambda_1)(z_i^2-1)}{M_i}\right) K(\frac{t_i-t}{h}) \\ \frac{\partial l}{\partial \sigma} &= \sum_i \frac{(z_i^2-1)}{(\sigma+(t_i-t)\sigma_1)} K(\frac{t_i-t}{h}) \\ \frac{\partial l}{\partial \sigma_1} &= \sum_i \frac{(z_i^2-1)}{(\sigma+(t_i-t)\sigma_1)} K(\frac{t_i-t}{h}) \end{aligned}$$

and are computed using 3 functions.

5.4.5 Applications

The above method is firstly applied to calculate serum quantiles in the set $\{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$ of the second of the data sets described in chap-

ter one. According to formula $\hat{q}_p(x) = M(x)(1+L(x)S(x)\Phi^{-1}(p))^{1/L(x)}$, the seven quantile curves for IgG data shown in Figure 5.1 are based on the L, M and S curves displayed in Figure 5.3 which are fitted using local constant kernel method with bandwidth $h = 0.54$ which is h_p of Chapter 2. Correspondingly, Figure 5.2 is drawn with the L, M and S curves in Figure 5.4. Because of local linear fitting, unlike Figure 5.3, Figure 5.4 also displays L_1, M_1 and S_1 curves which smoothly estimate the derivative curves of L, M and S according to the theory of local linear fitting. The L, M and S curves in Figure 5.3 are used to initialise estimator in forming Figure 5.4 with 10 iterations.

Comparing Figure 5.1 and 5.2, both left and right tails of quantile curves are a bit less spread out than those of Figure 5.2, but there is not much difference on median parts. Similarly, Figure 5.5 is seven quantile curves for U.S girl data from set $\{0.03, 0.1, 0.25, 0.5, 0.75, 0.9, 0.97\}$, which are derived from the L, M and S curves in Figure 5.7 using local constant fiting with $h = 1.8$, while Figure 5.6 and 5.8 are the local linear fitting versions of Figure 5.5 and 5.7, respectively.

Comparing Figures 5.5 and 5.6, local linear fitting the body weight data for age less than 3 is much better than local constant fitting, however, comparing either Figure 5.5 or Figure 5.6 to Figure 3.6 of Chapter 3, in terms of fitting the quantile curves in the low age area of U.S data, this semi-parametric method is better than some nonparametric methods, since it can avoid quantile crossing in this special area as mentioned generally in Section 5.1.

Further, Figures 5.9 to 5.12 are to fit Gambian data from set $\{0.03, 0.1, 0.25, 0.5, 0.75, 0.9, 0.97\}$ in terms of local constant and local linear fitting with bandwidth $h = 2.5$. Comparing Figure 5.9 and 5.10 with Figure 3.3 in Chapter 3, the advantage of semi-parametric method over nonparametric method

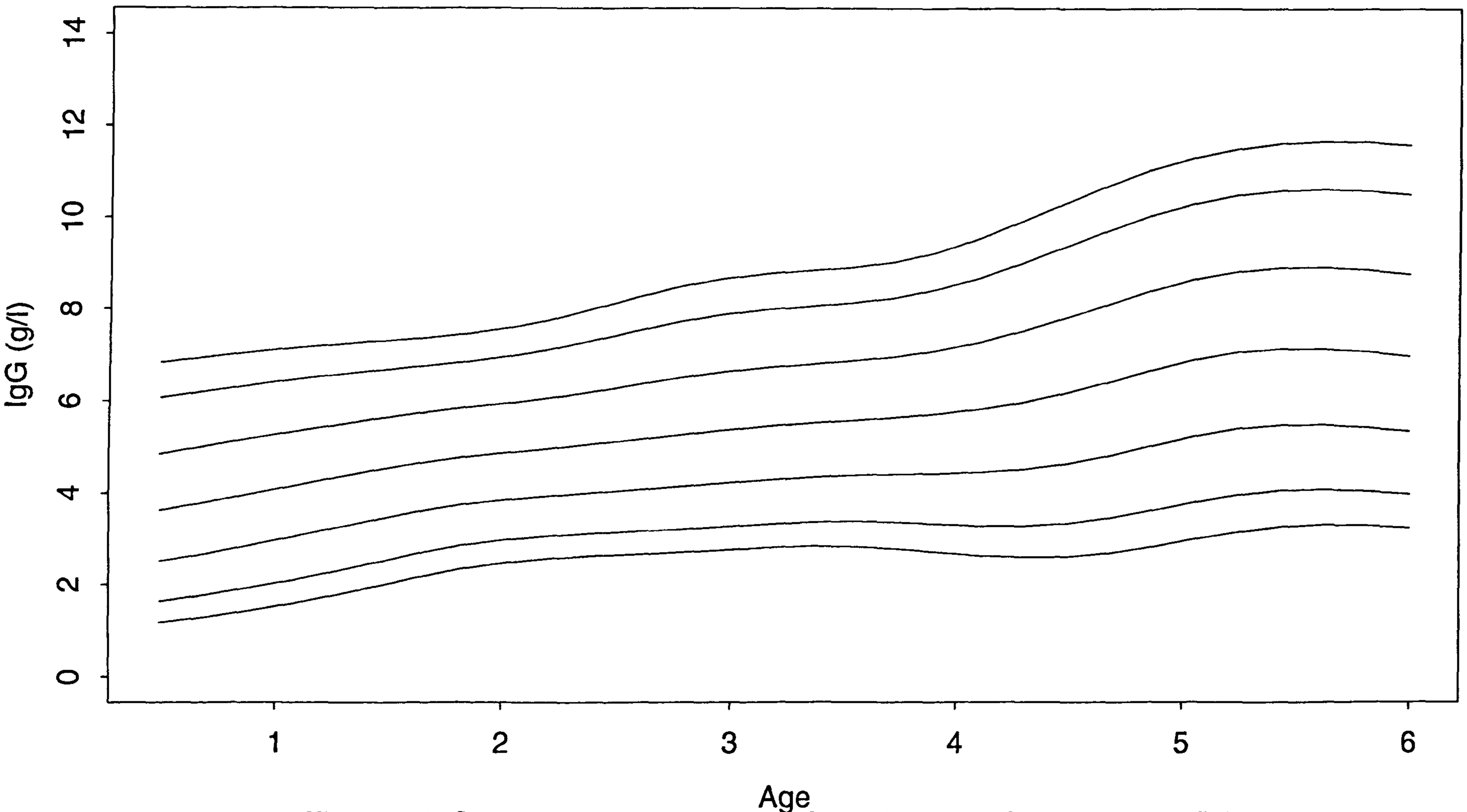


Figure 5.1: Seven quantiles smoothed for IgG data by local constant fitting

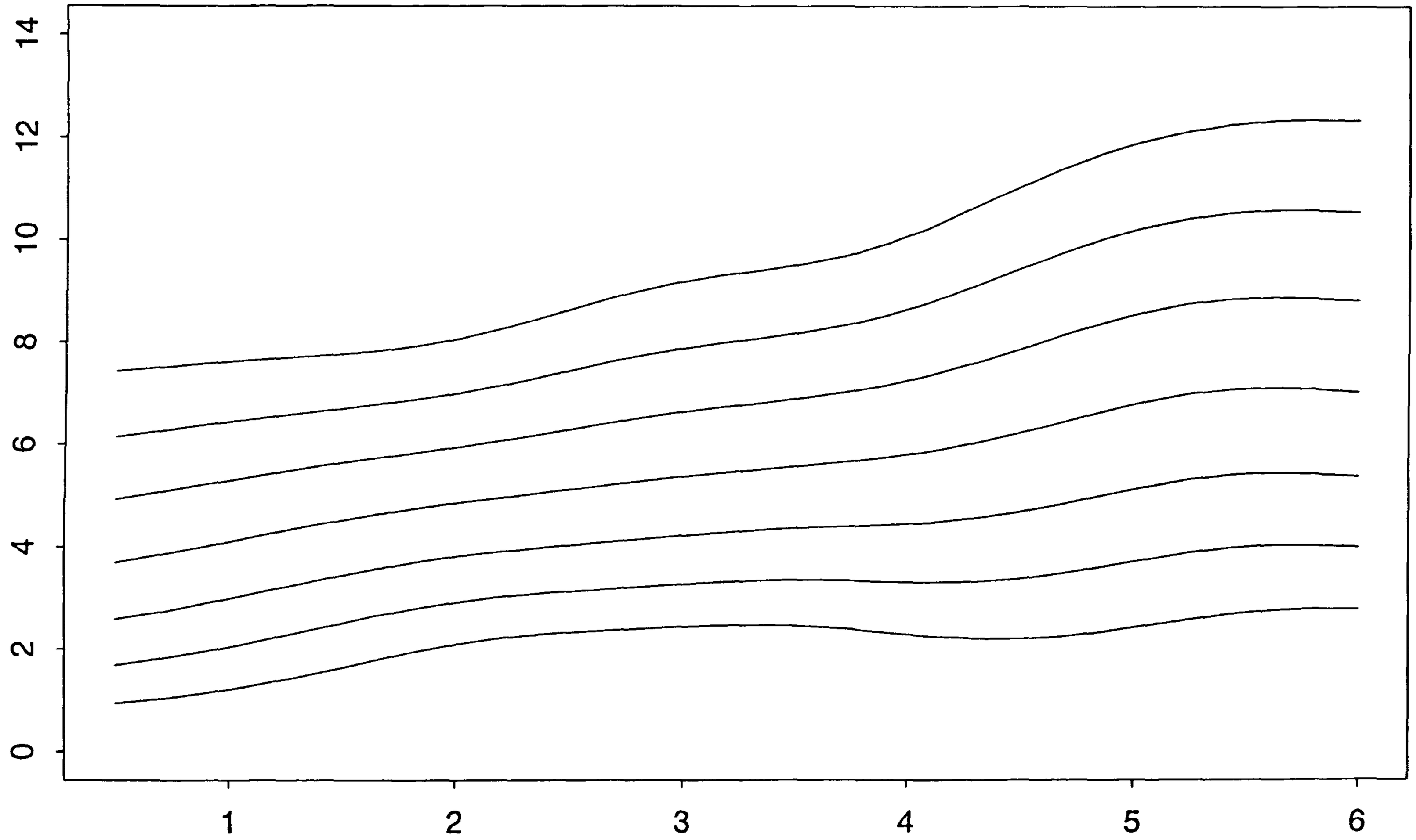


Figure 5.2: Seven quantiles smoothed for IgG data by local linear kernel version of LMS method

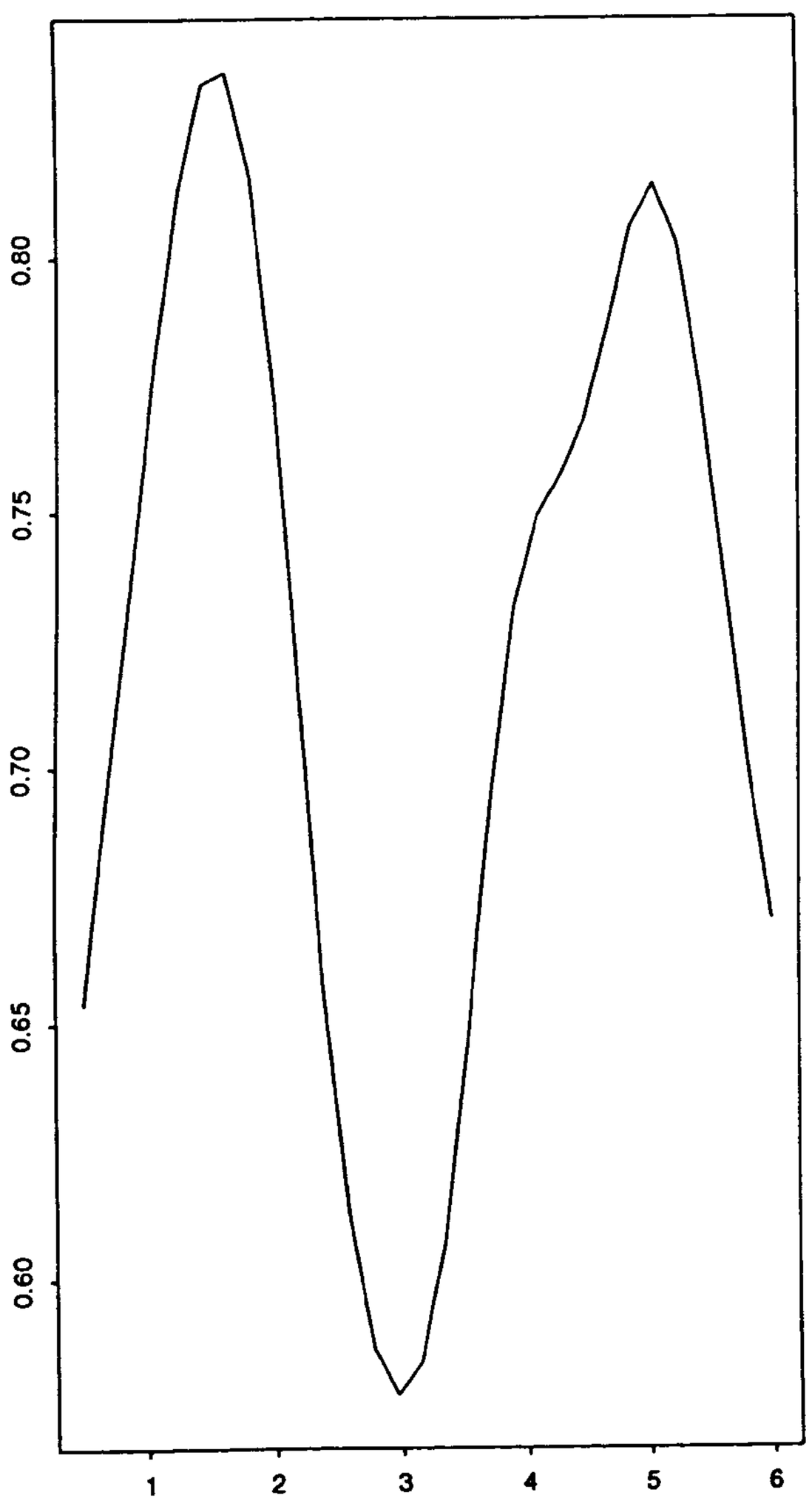
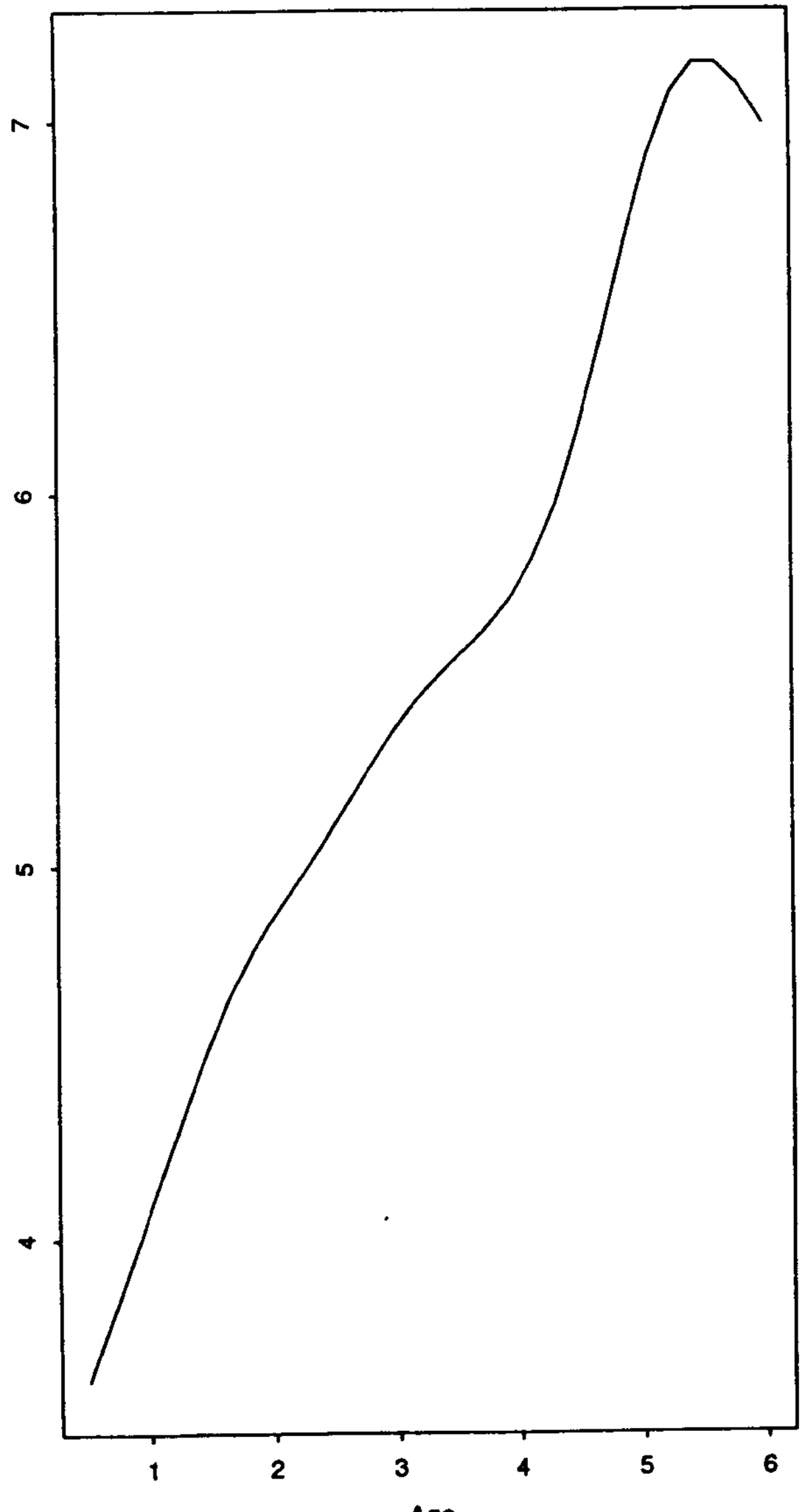
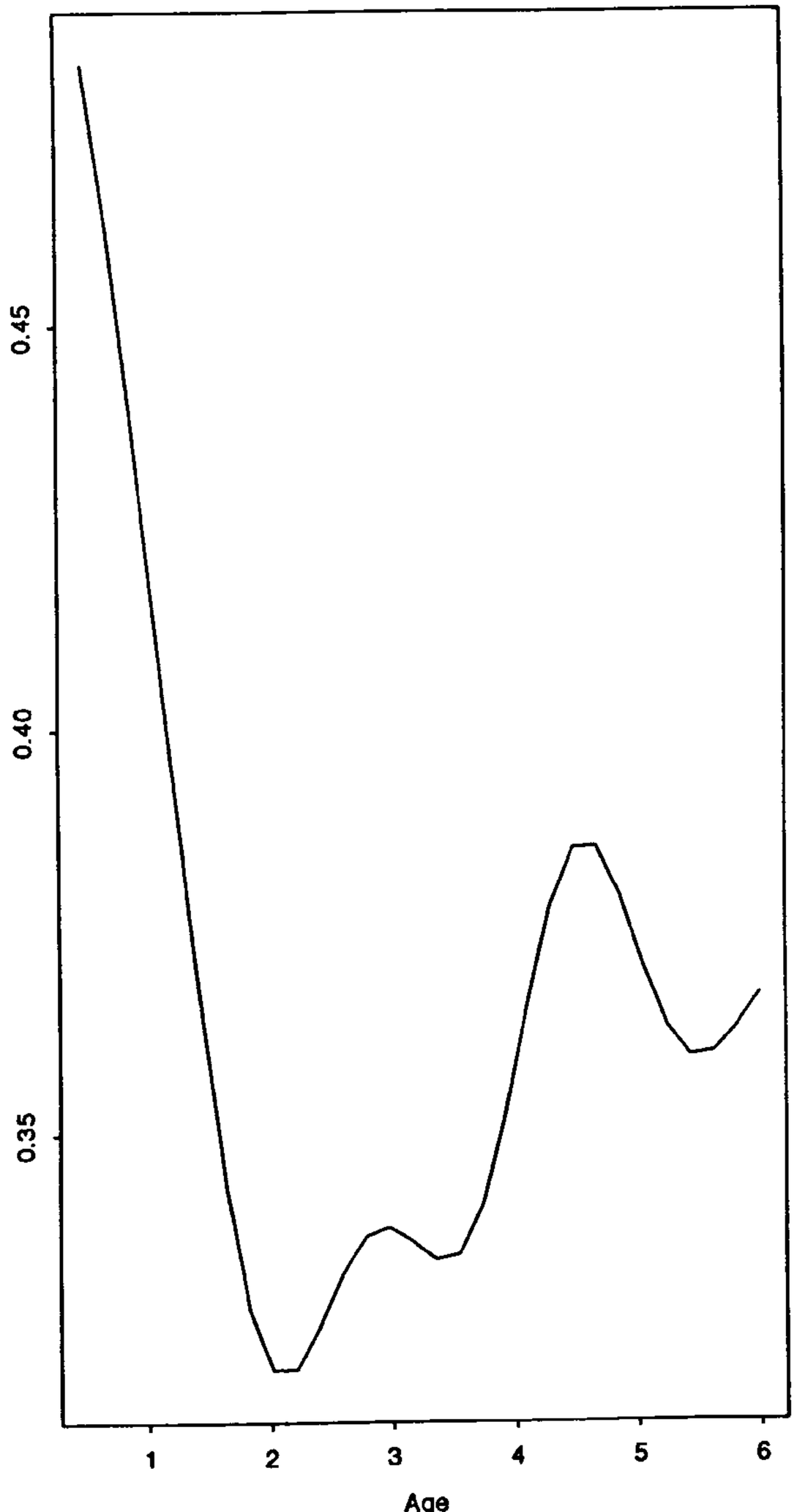
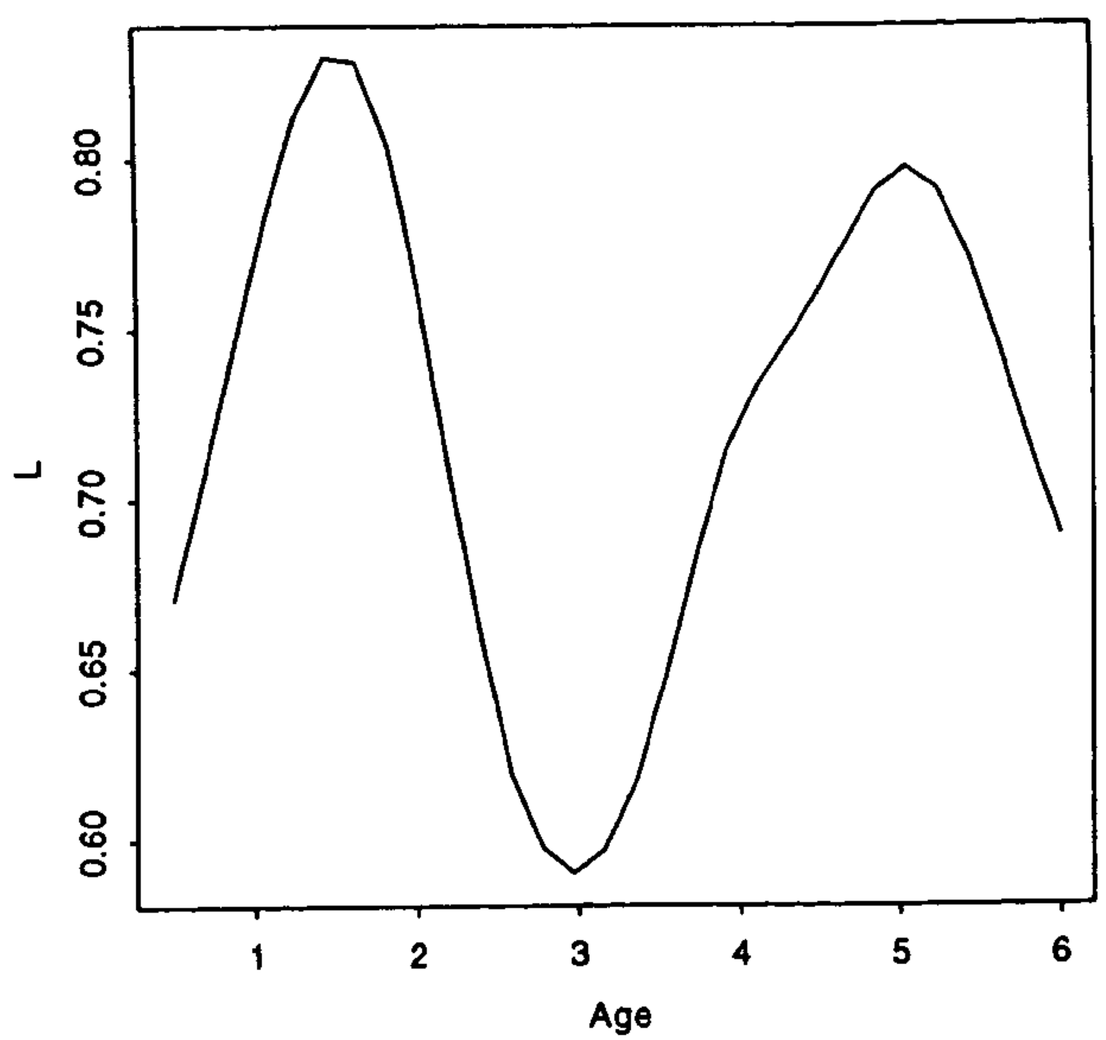
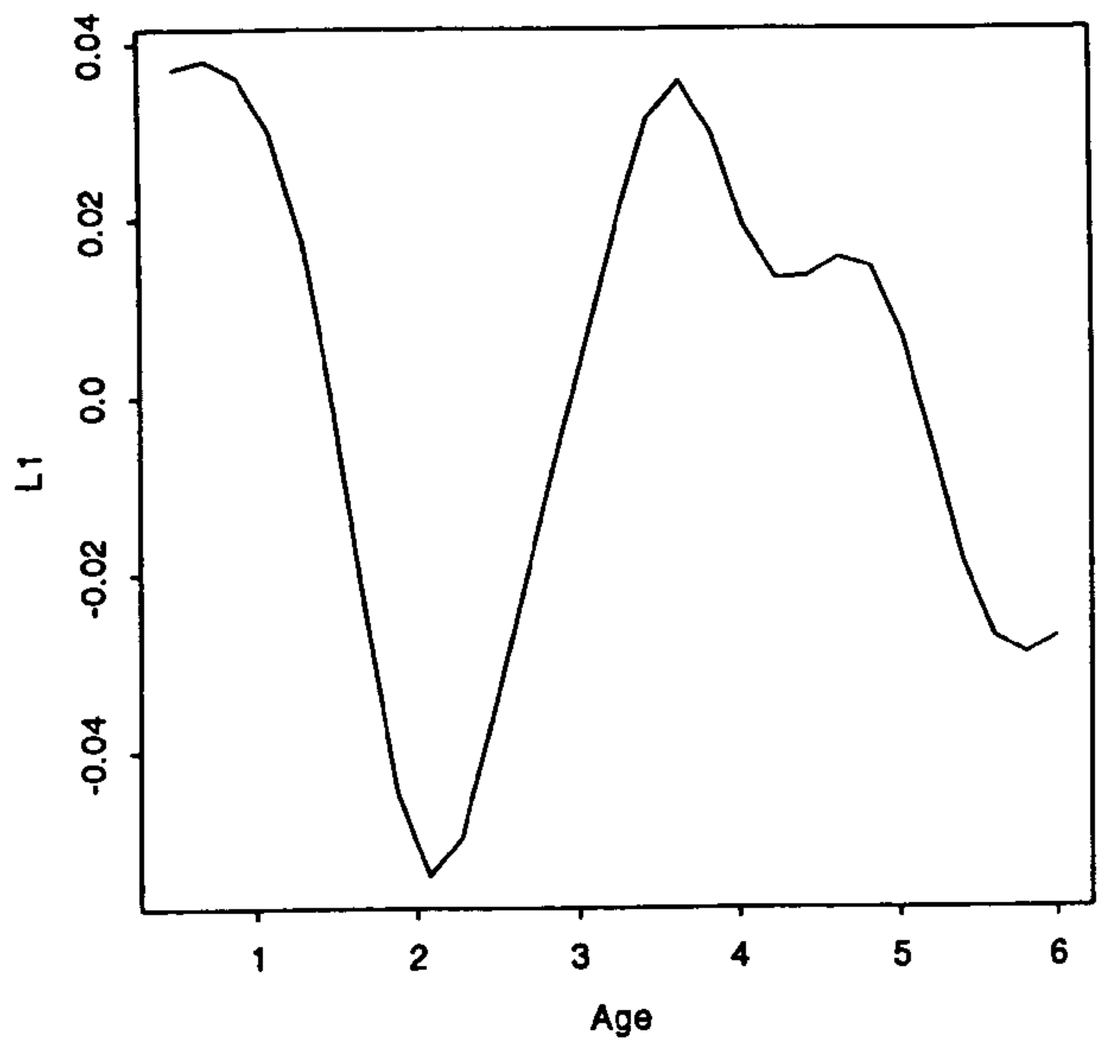


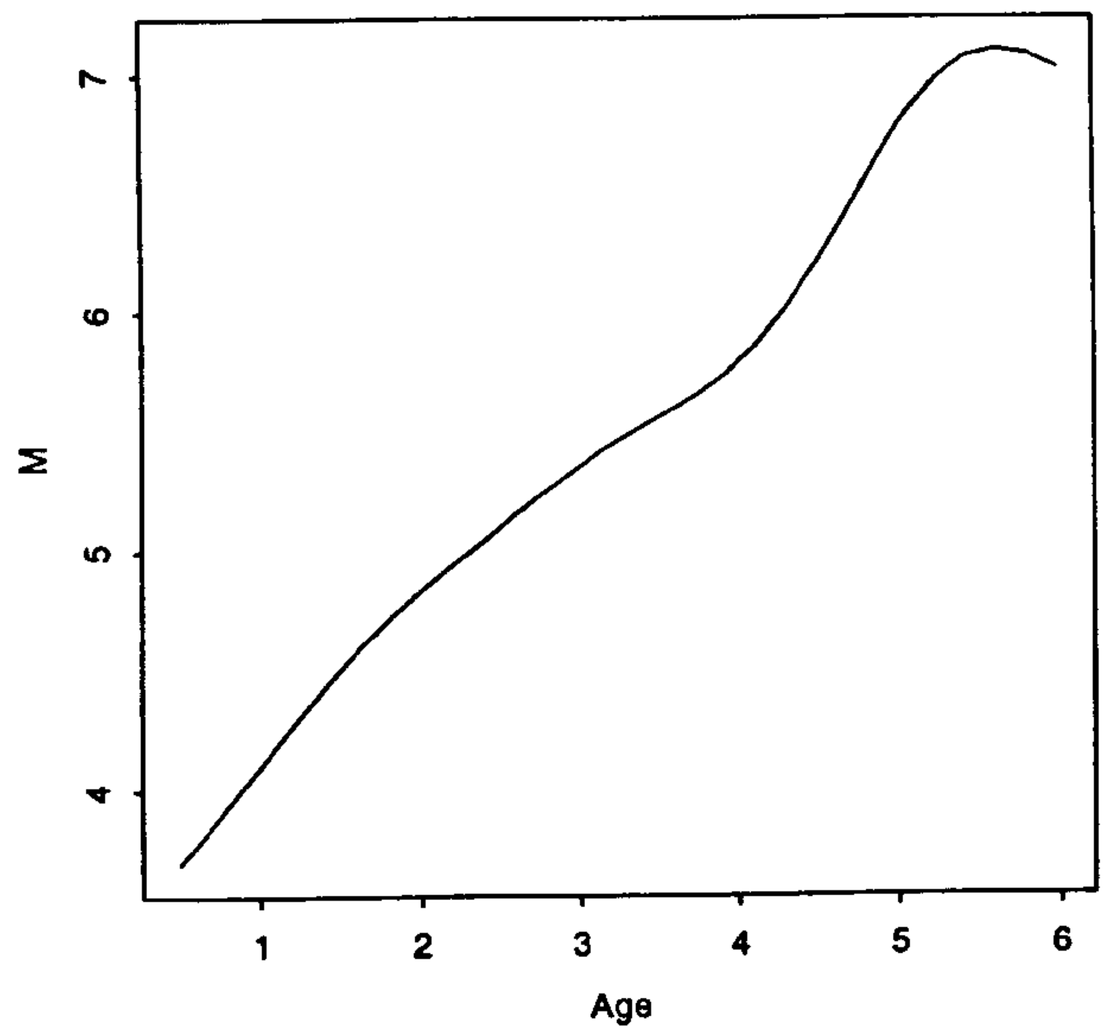
Figure 5.3: Box-Cox power for IgG data by local constant fitting

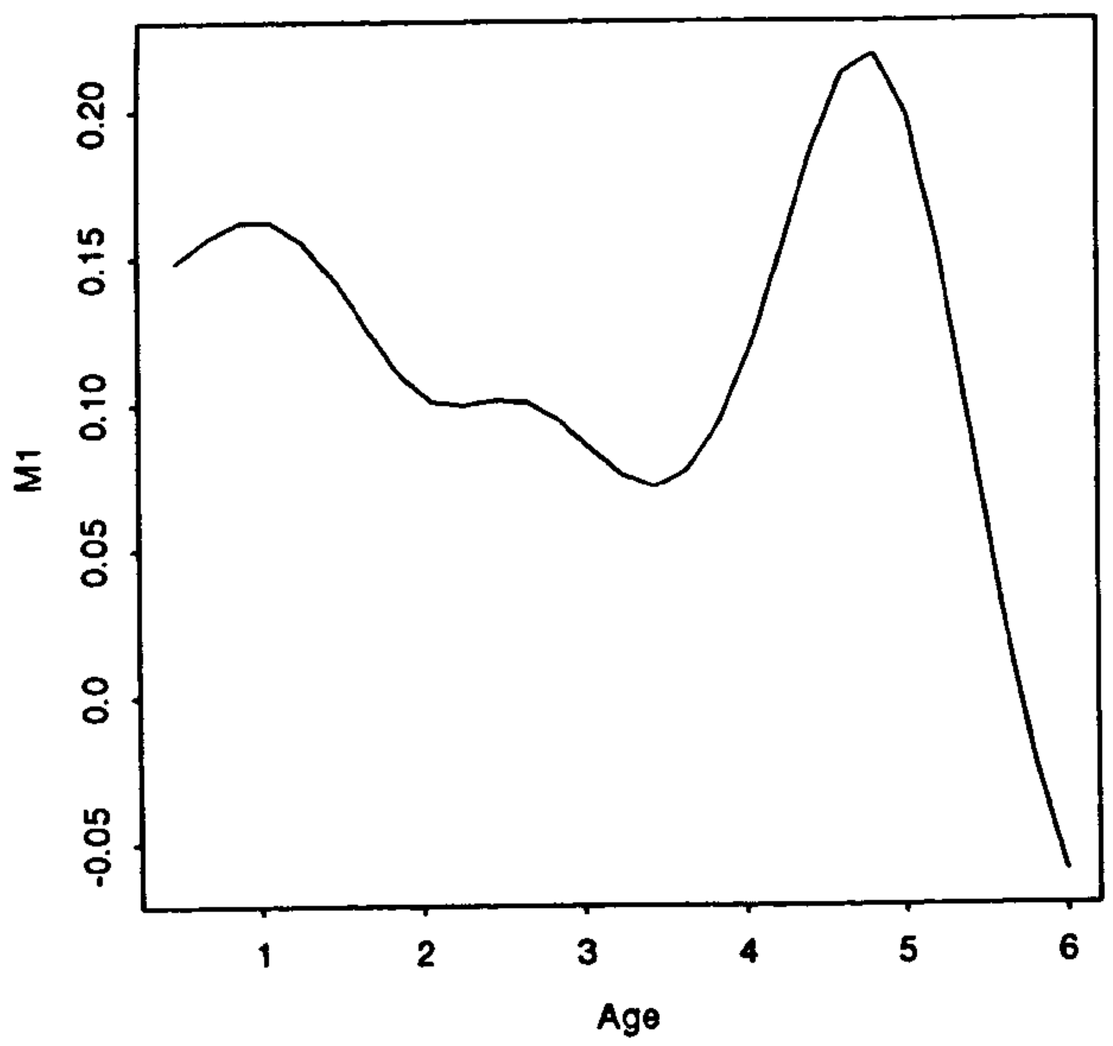


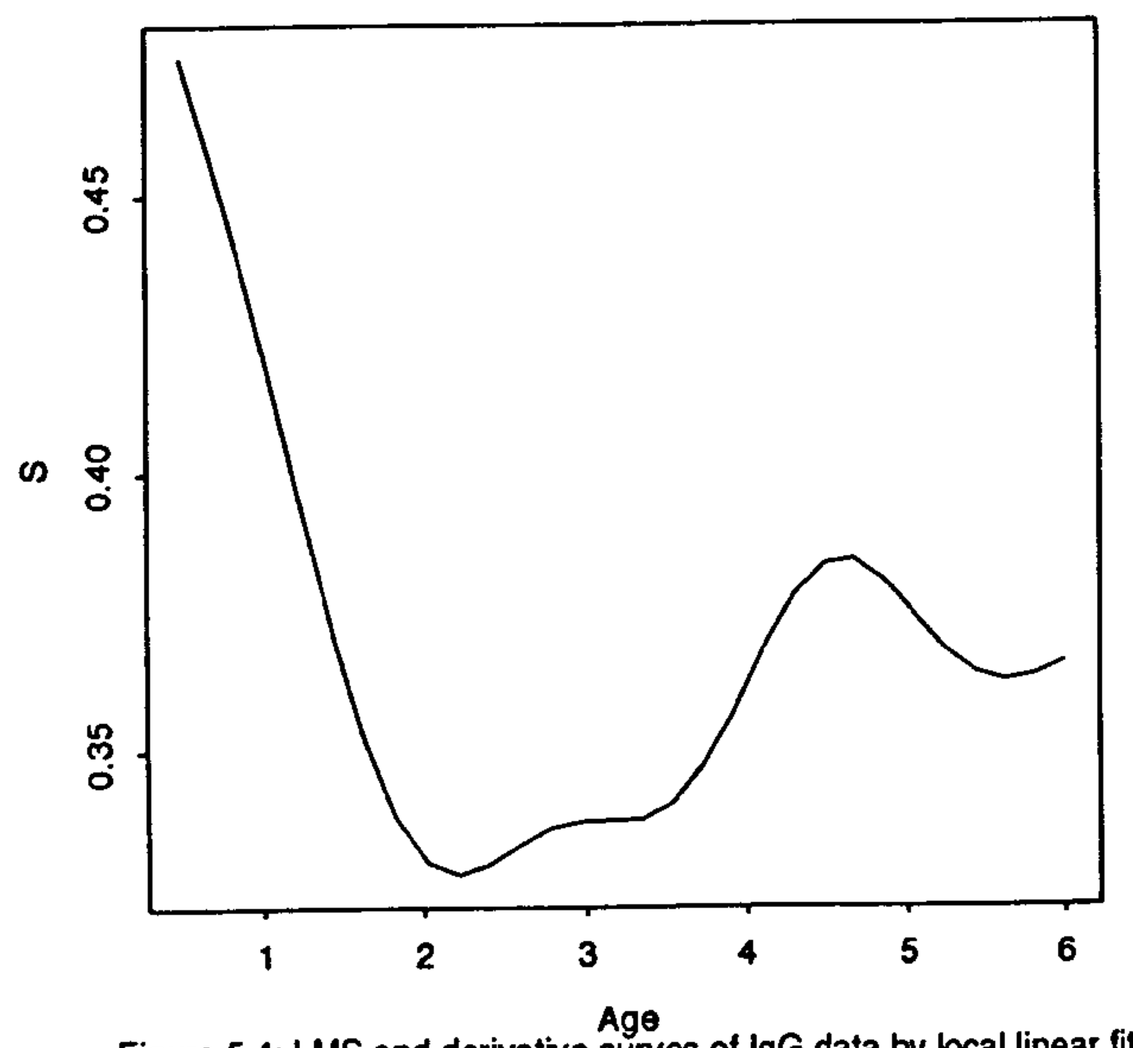












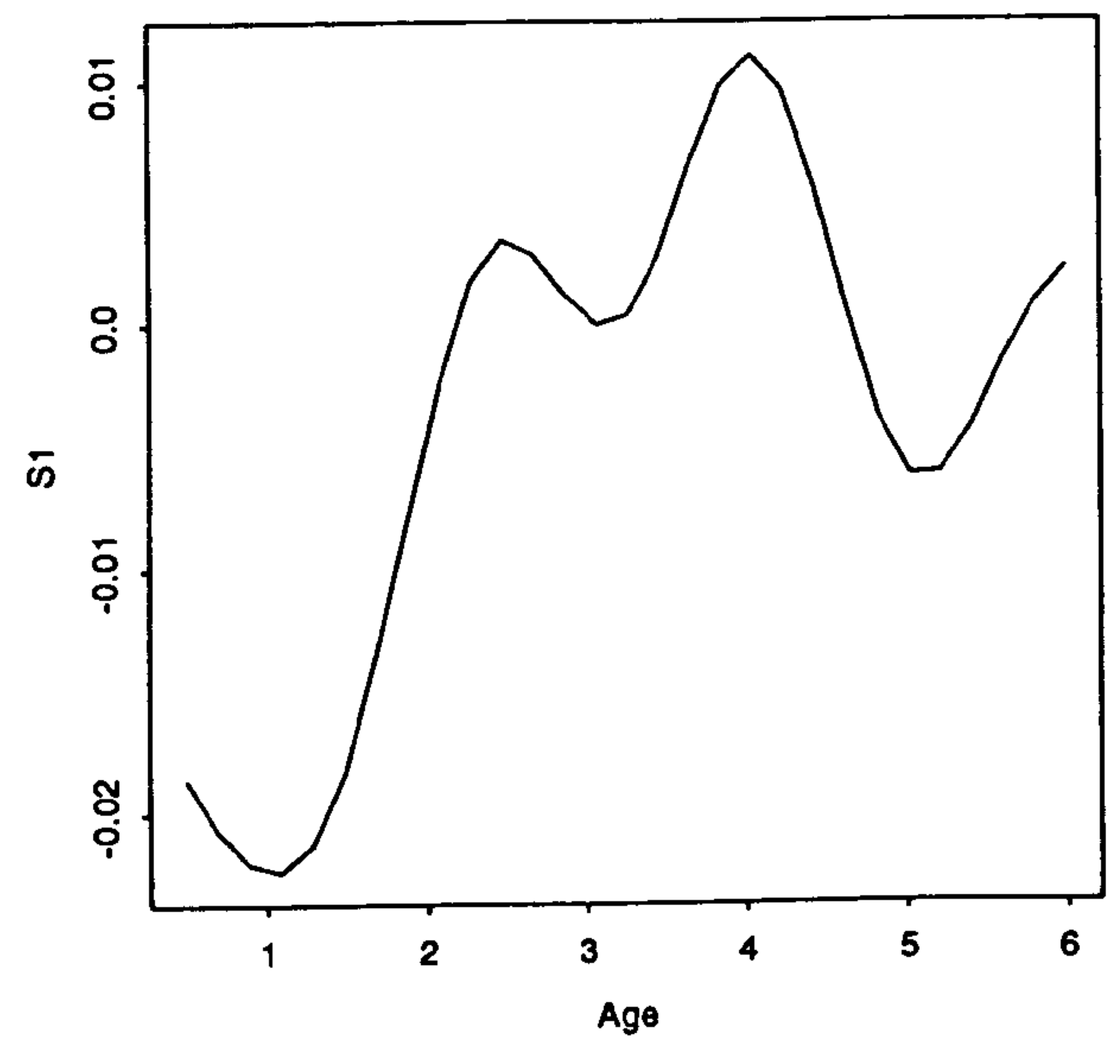


Figure 5.4: LMS and derivative curves of IgG data by local linear fit

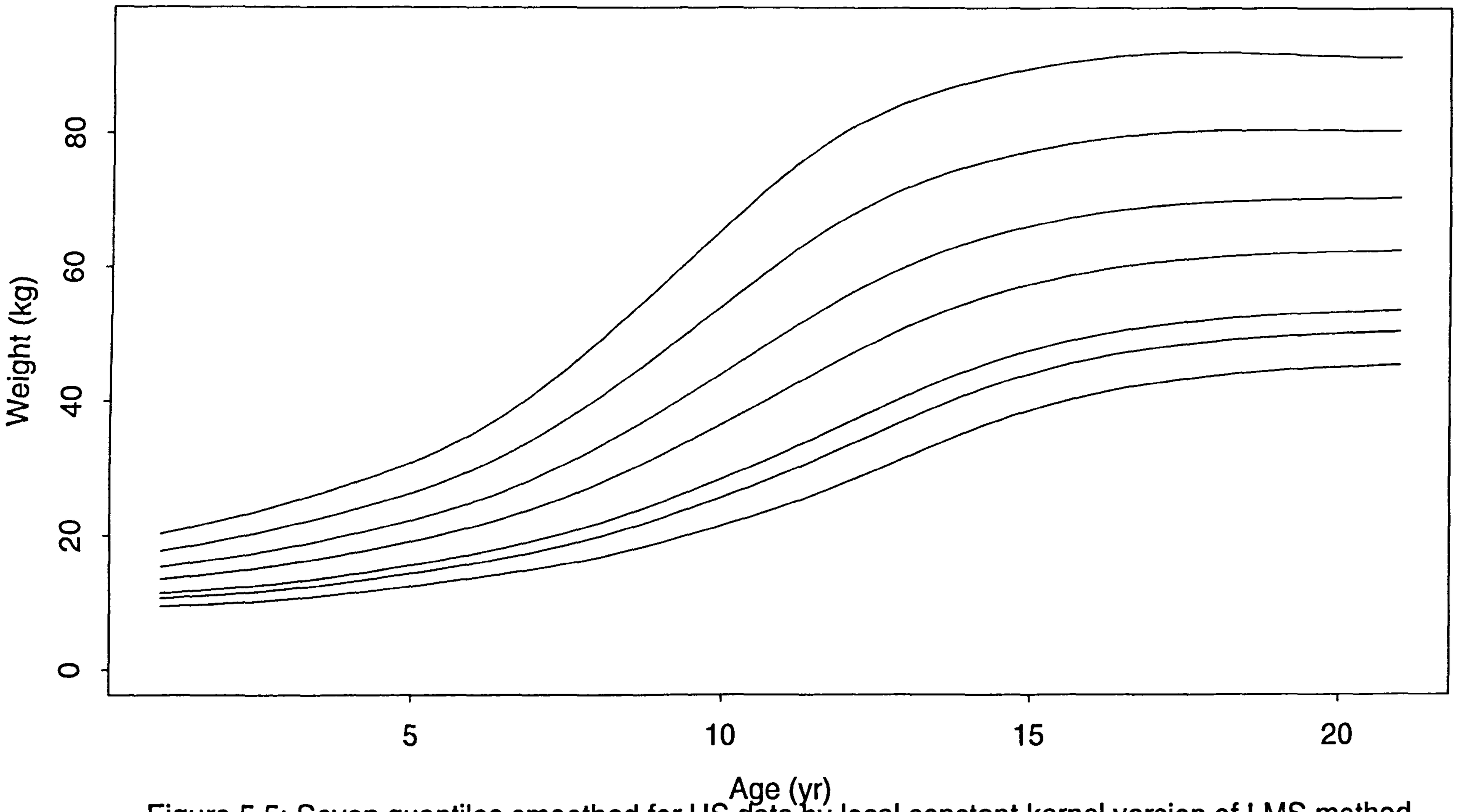


Figure 5.5: Seven quantiles smoothed for US data by local constant kernel version of LMS method

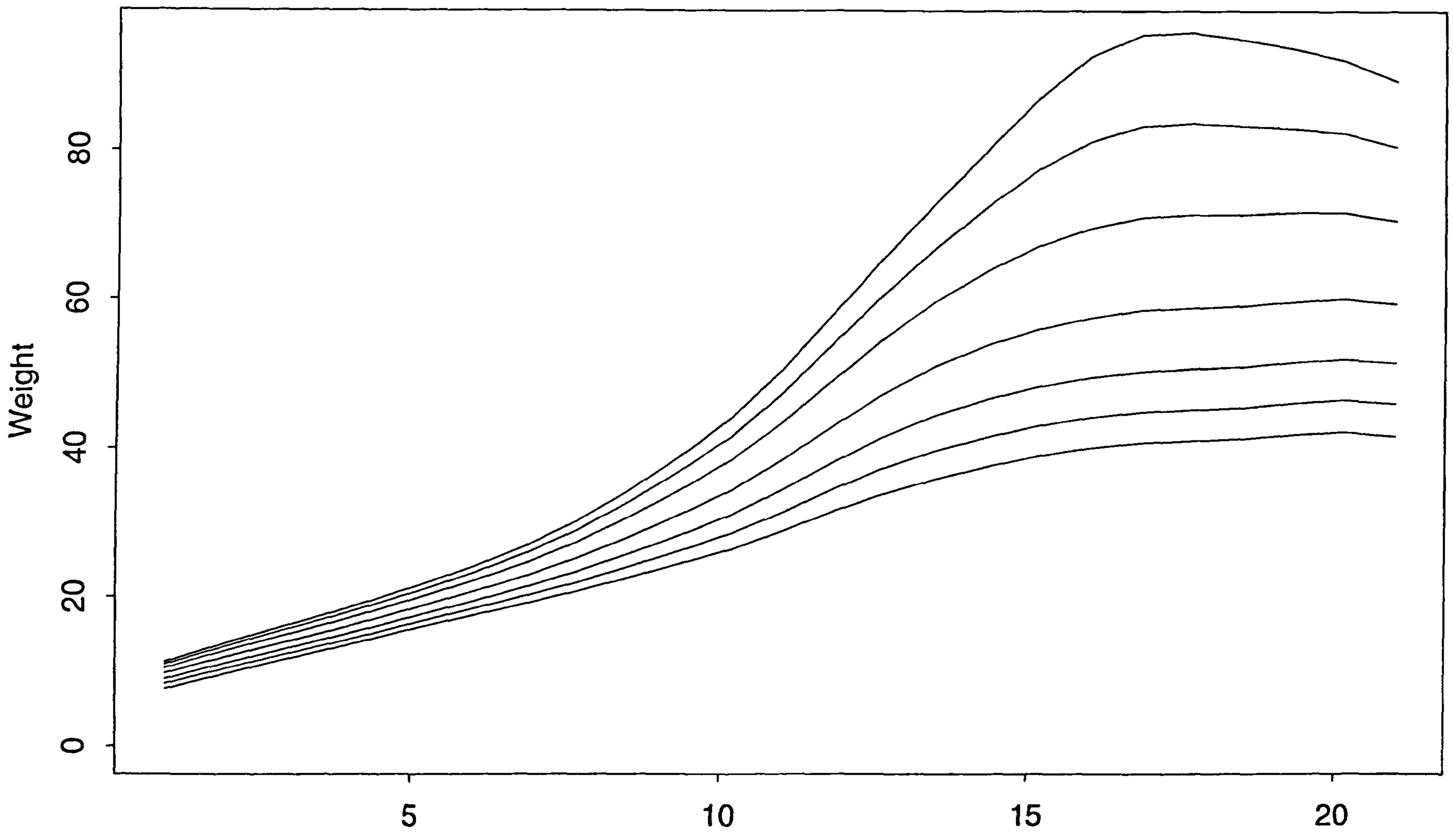
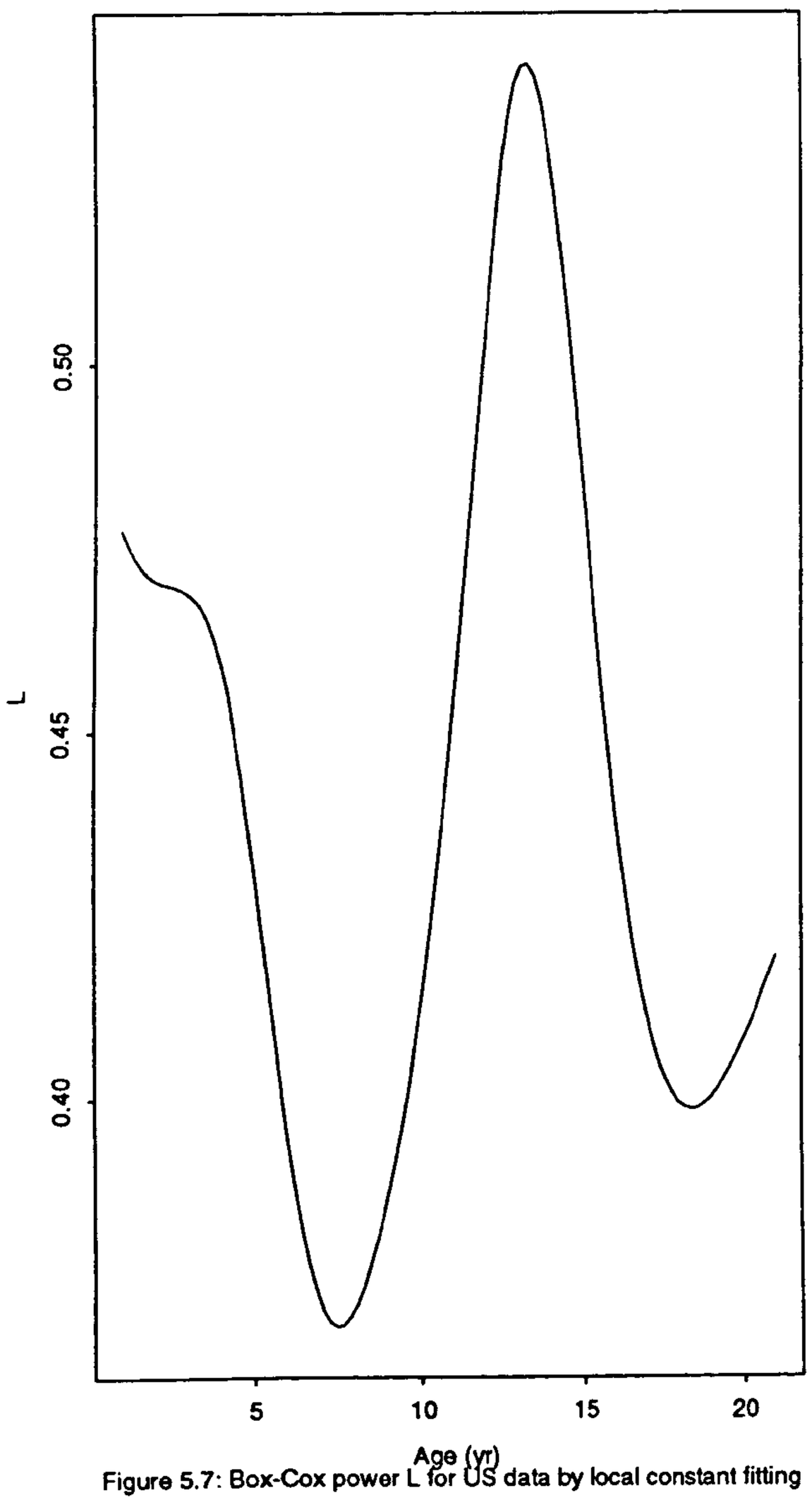
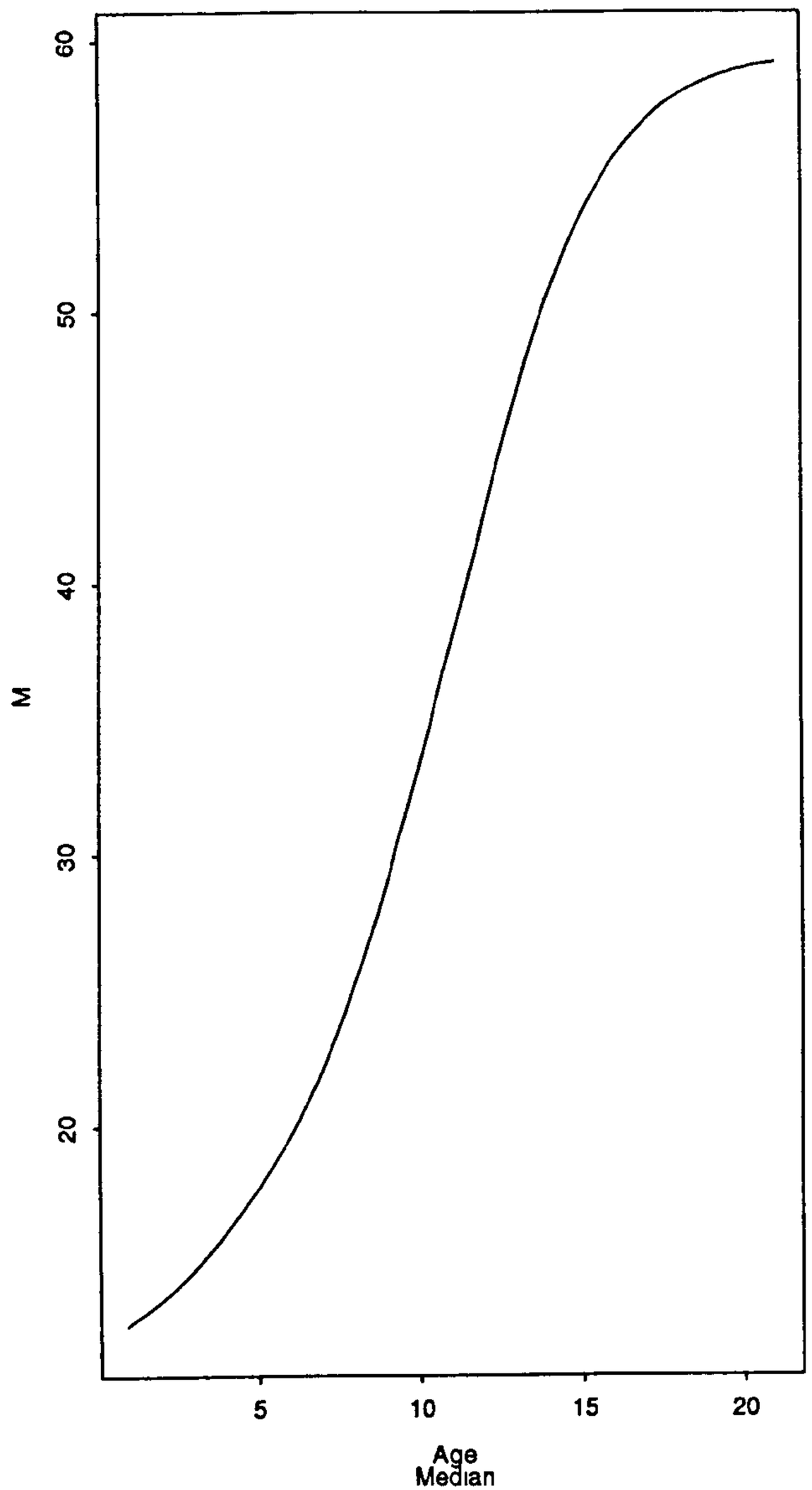


Figure 5.6: Seven quantiles smoothed for US data by local linear kernel version of LMS method





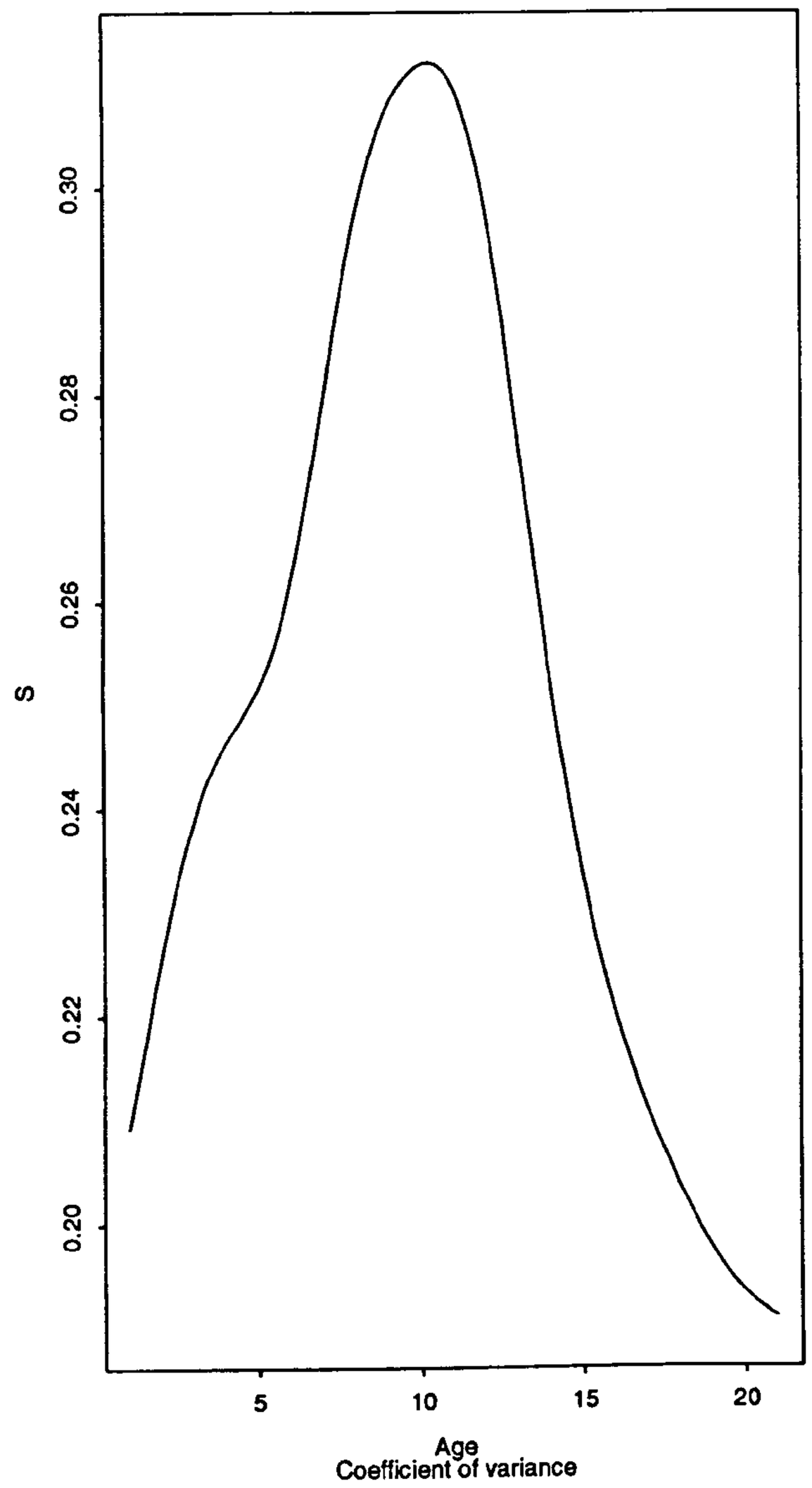
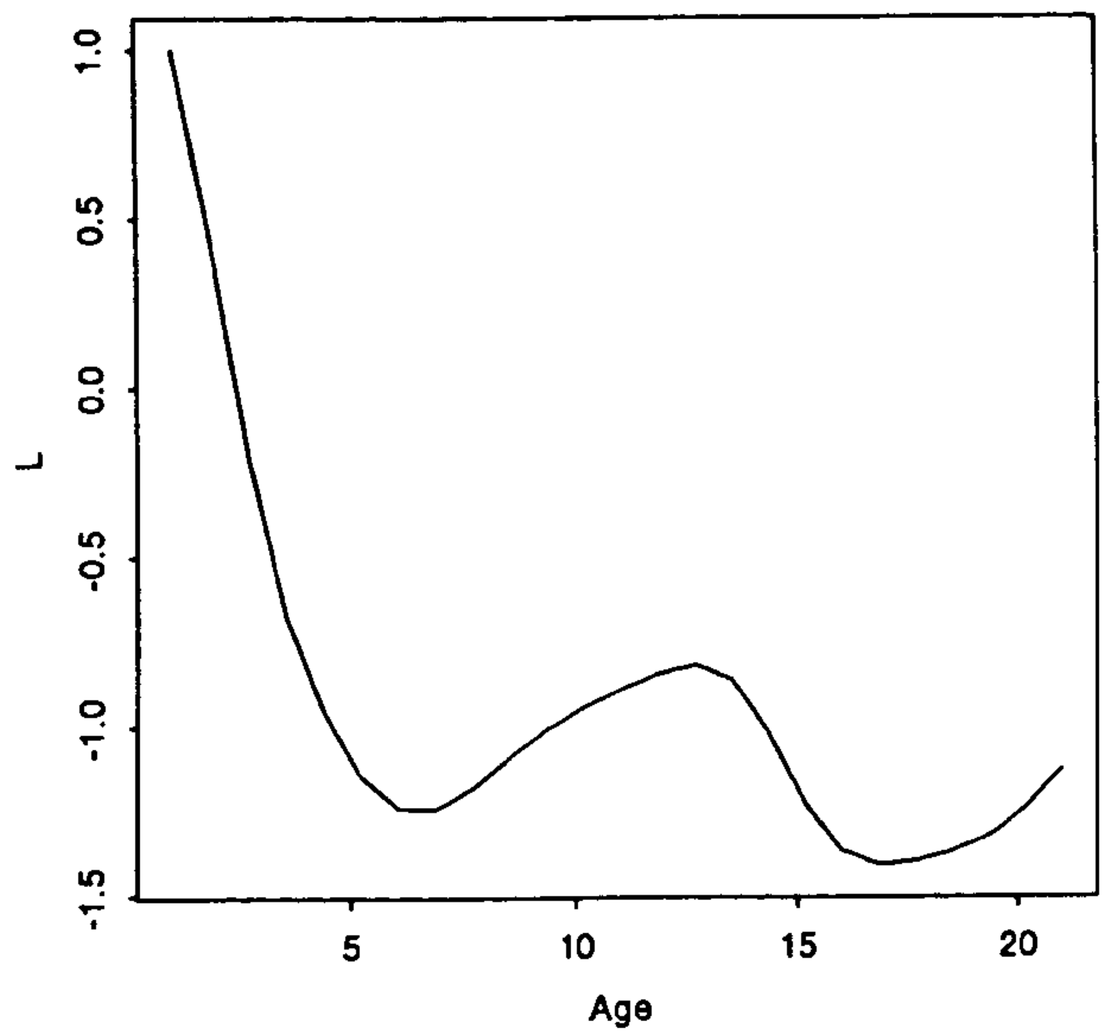
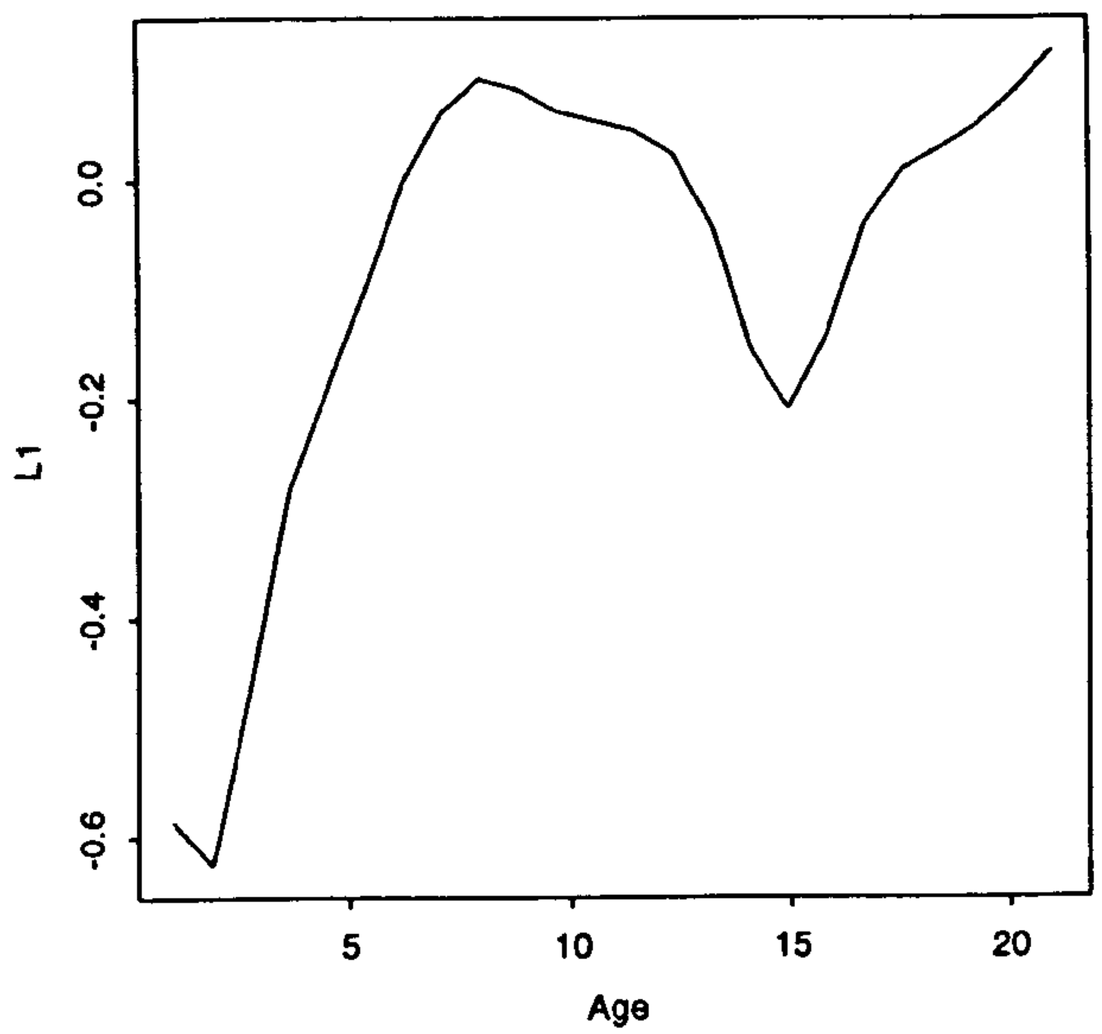
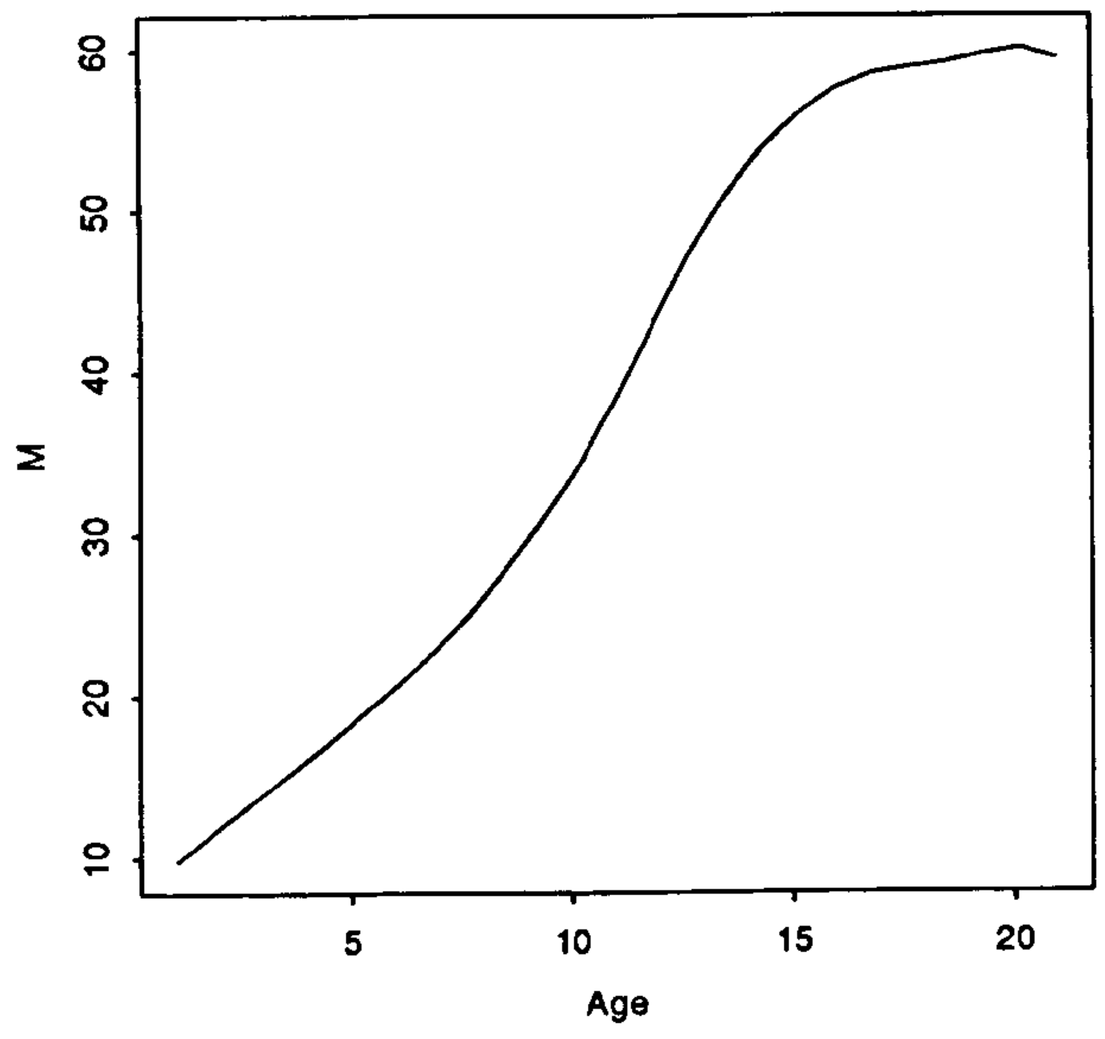
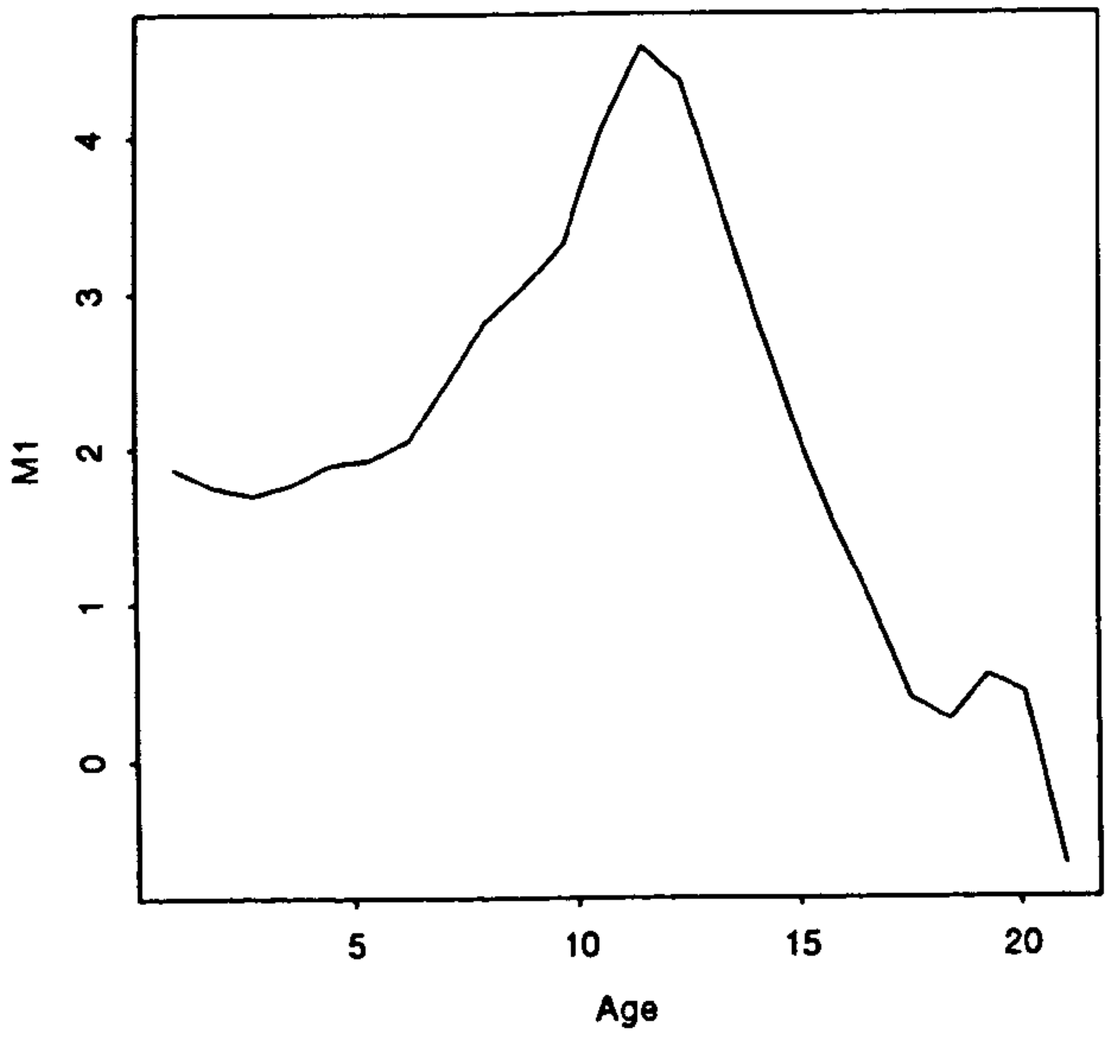


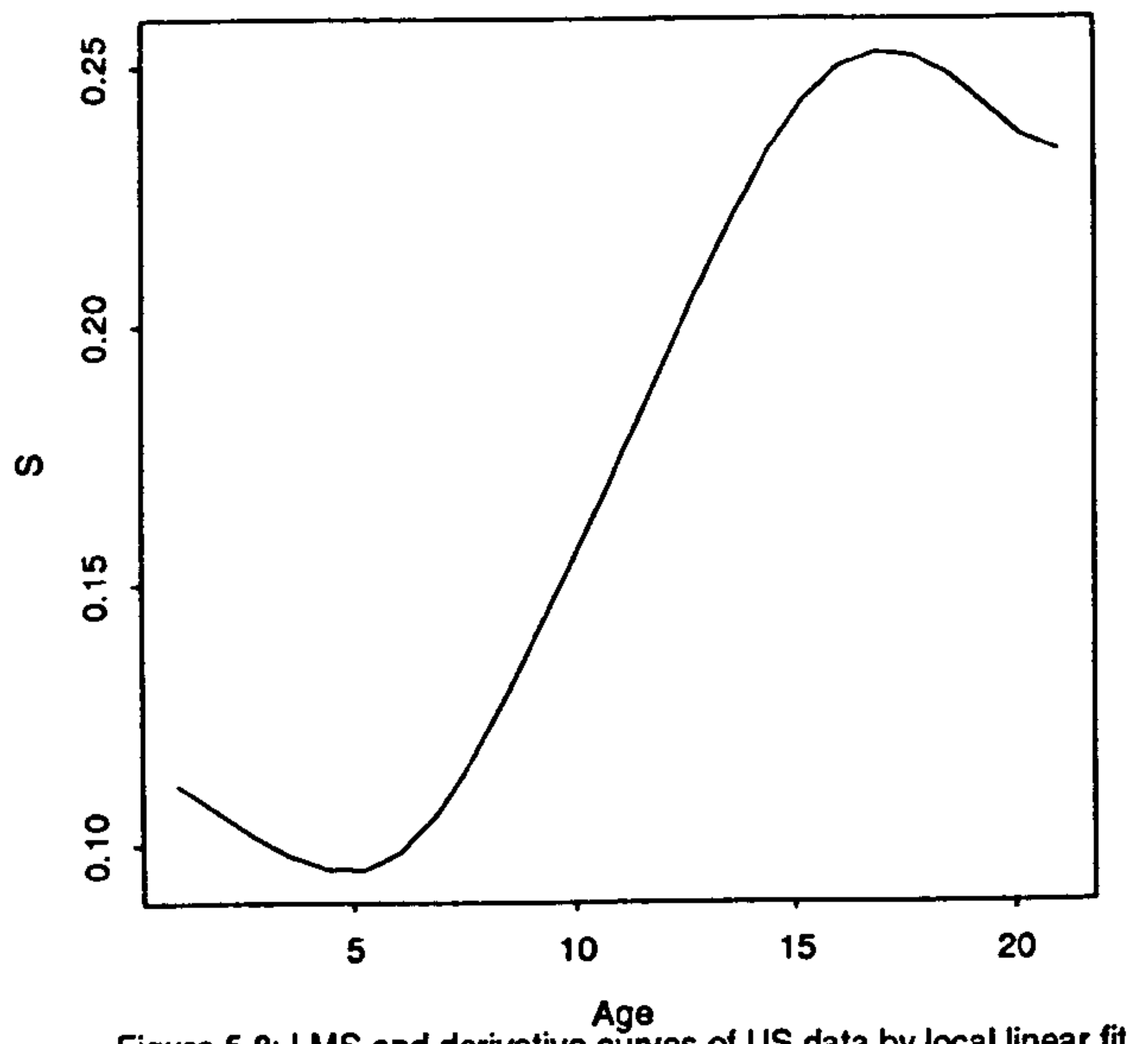
Figure 5.7: Box-Cox power L for US data by local constant fitting











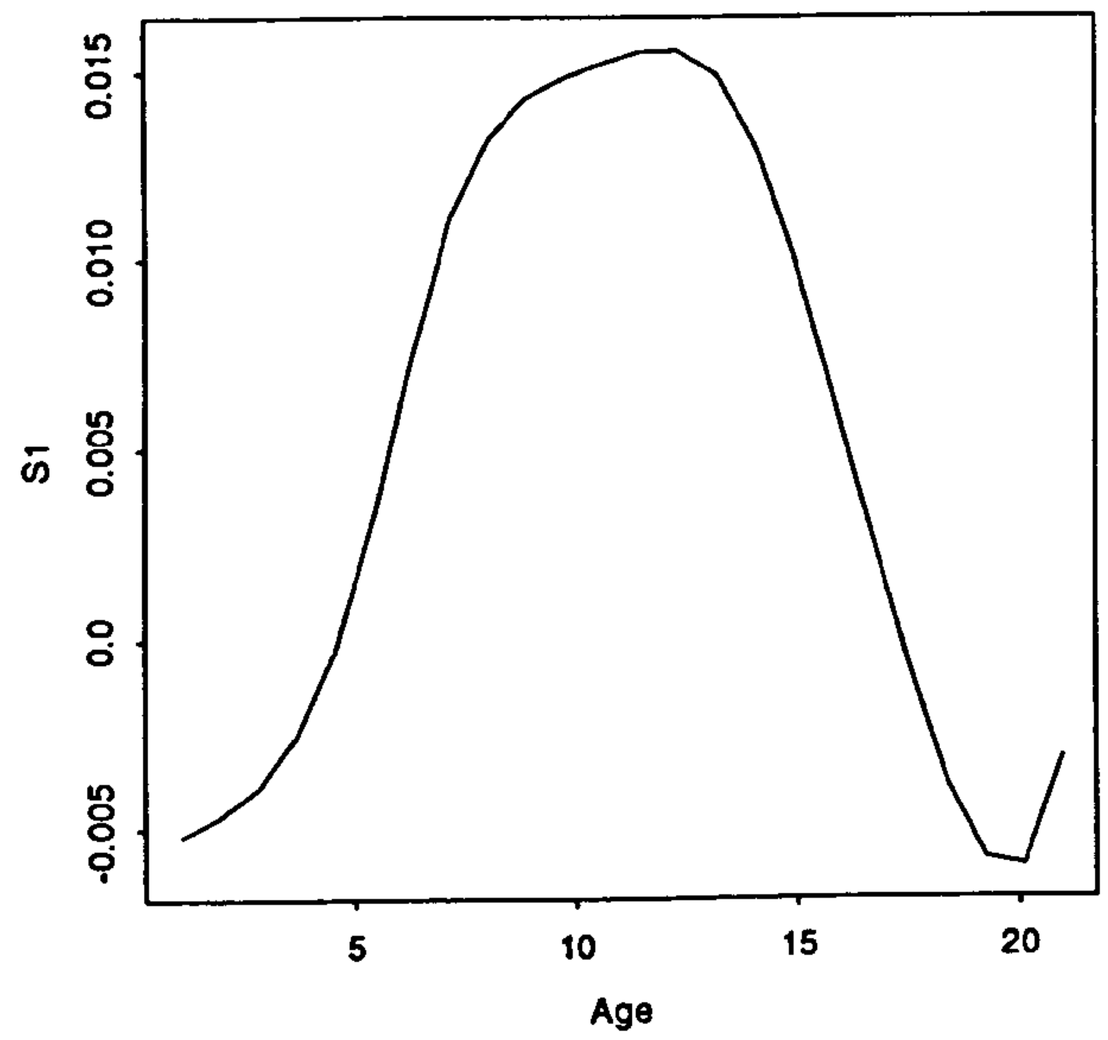


Figure 5.8: LMS and derivative curves of US data by local linear fit

for this data appears on fitting the notch position of Gambian data .

Comparing Figures 5.3 and 5.4, the three curves L, M and S fitted by both local constant and local linear methods are almost identical, so are Figure 5.1 and 5.2. This should be attributed to the approximate normality of IgG data. Comparing Figures 5.7 and 5.8 or Figures 5.11 and 5.12, however, both transformation power curve L and the coefficient of variation curve S by two fittings are different in spite of no change on median curve M. Particularly, L curves by local linear fitting fall well below zero, this may result from both US weight data and Gambian Triceps skinfold data being considerably skew. This seems to suggest some lack of identifiability in fitting power and coefficient of variation. We prefer the curves given by local linear fitting.

In general the quantile curves using local linear fitting are always smooth enough although fitting of L, M and S's derivative is comparatively not satisfactory and usually require a little larger bandwidth.

We have not done anything sophisticated about bandwidth choice although we hope to have a rule-of-thumb as the one in Chapter 2 for this kind of kernel-based semi-parametric method, as the $MSE(\hat{q}_p(x))$ in Theorem 5.6 is too complicated to treat easily.

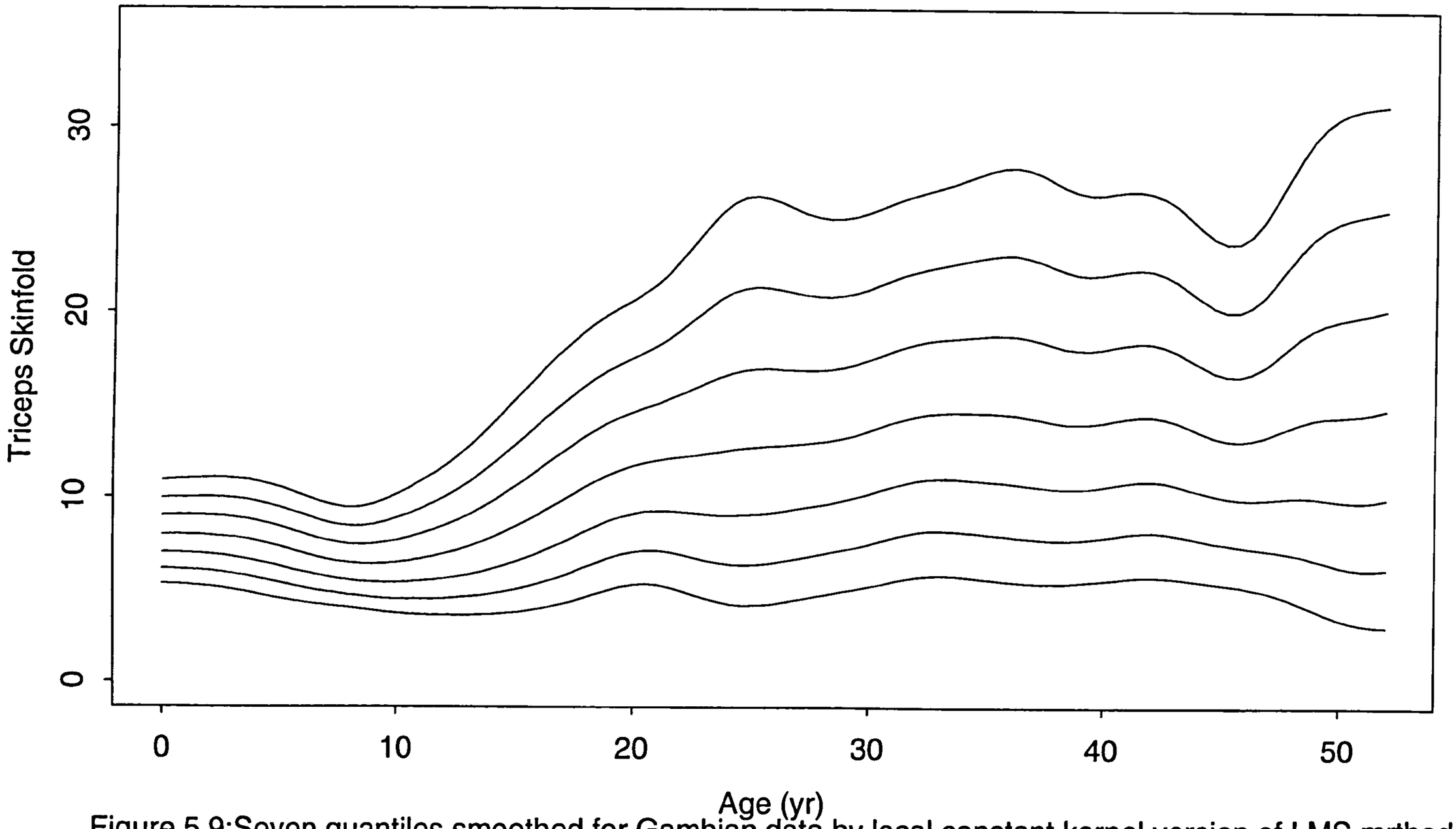


Figure 5.9:Seven quantiles smoothed for Gambian data by local constant kernel version of LMS mrthod

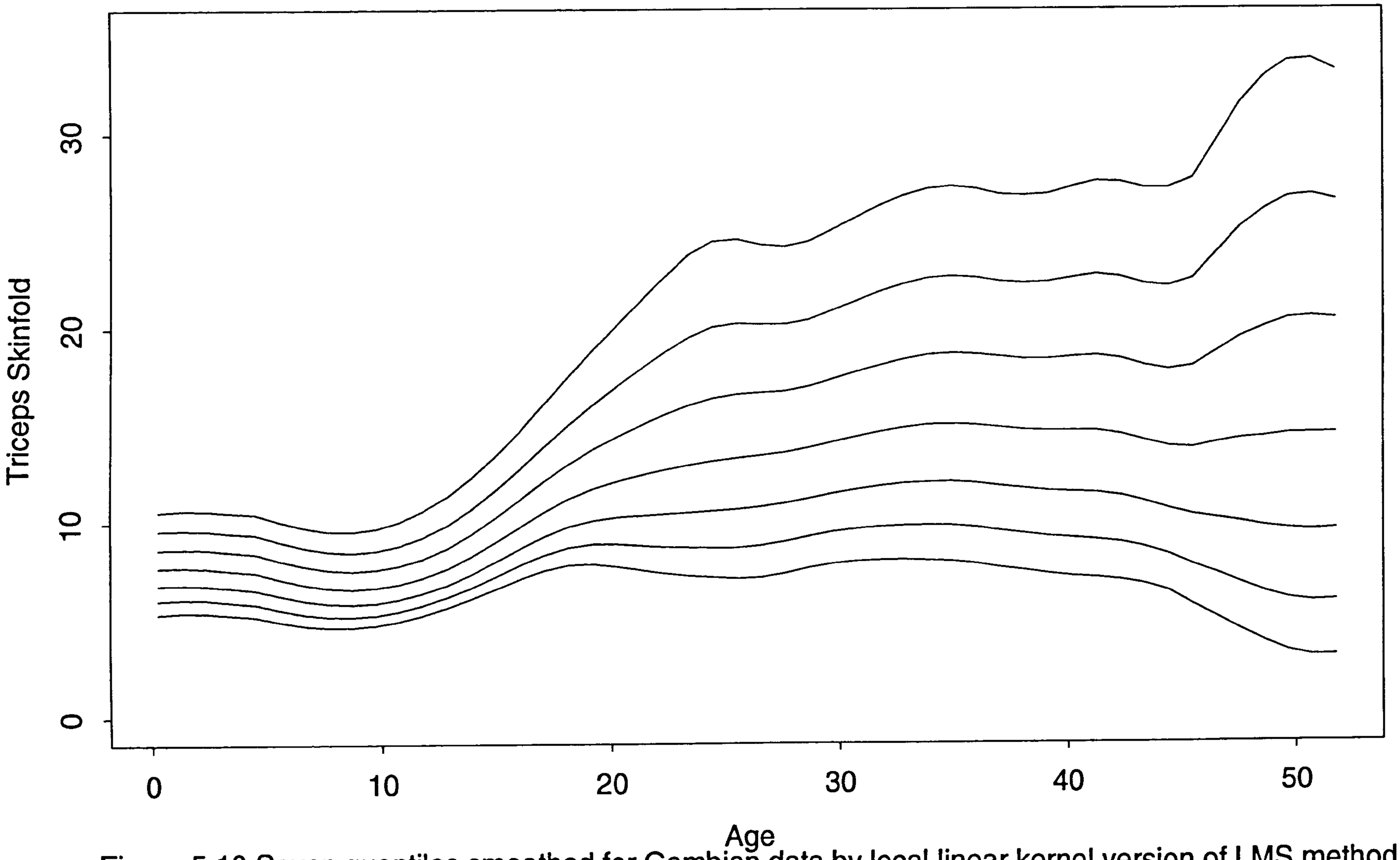


Figure 5.10:Seven quantiles smoothed for Gambian data by local linear kernel version of LMS method

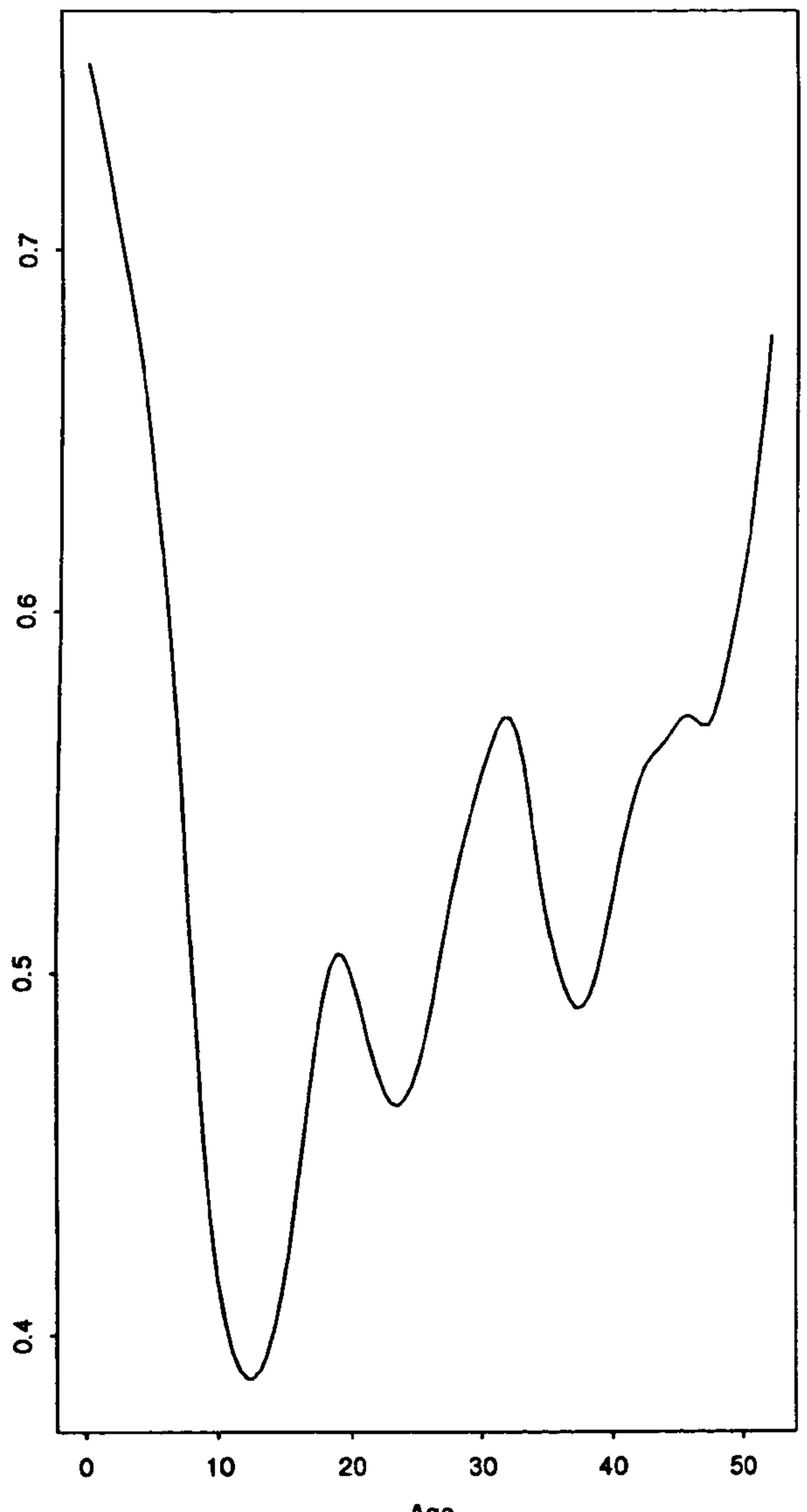
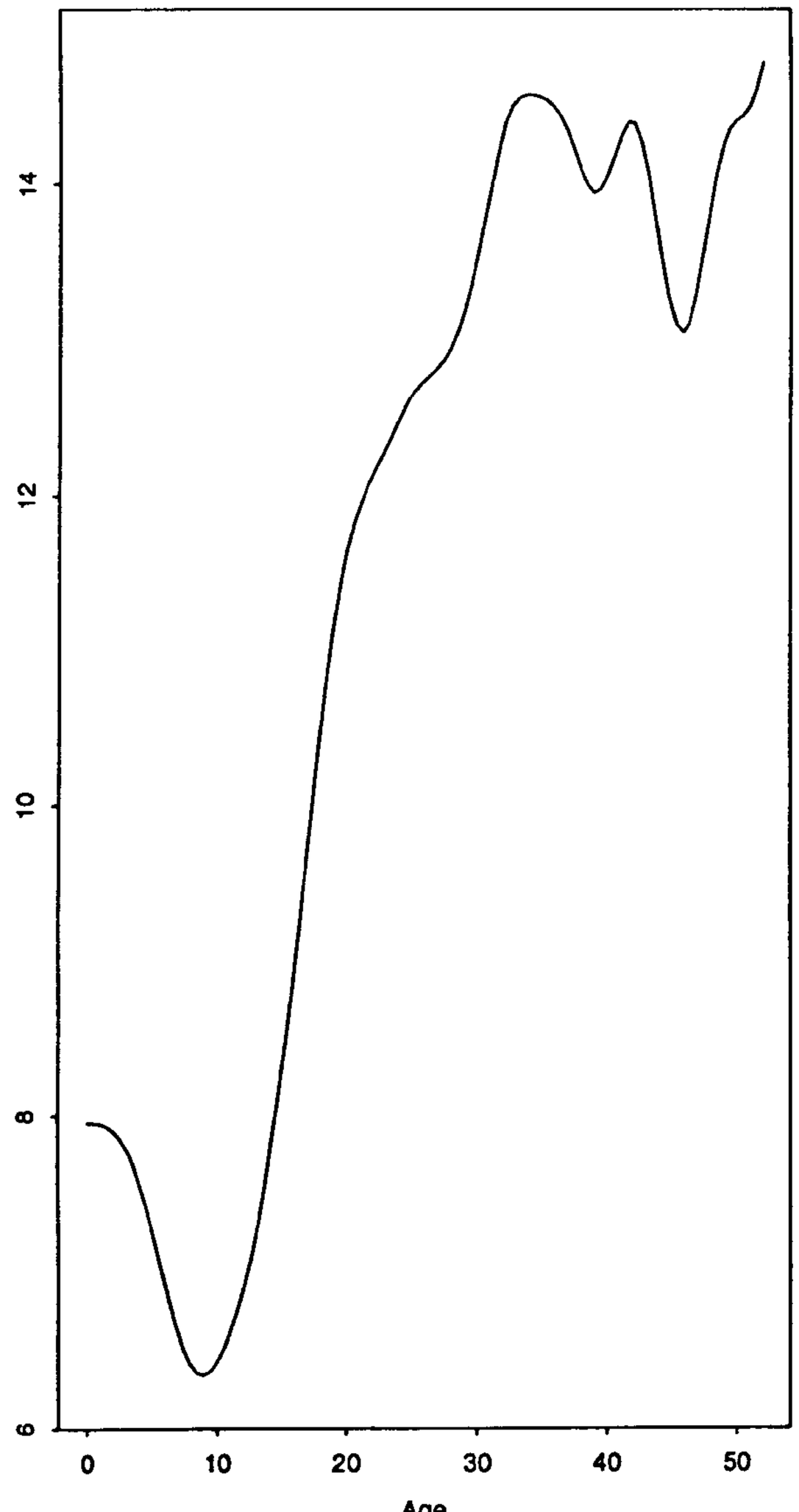
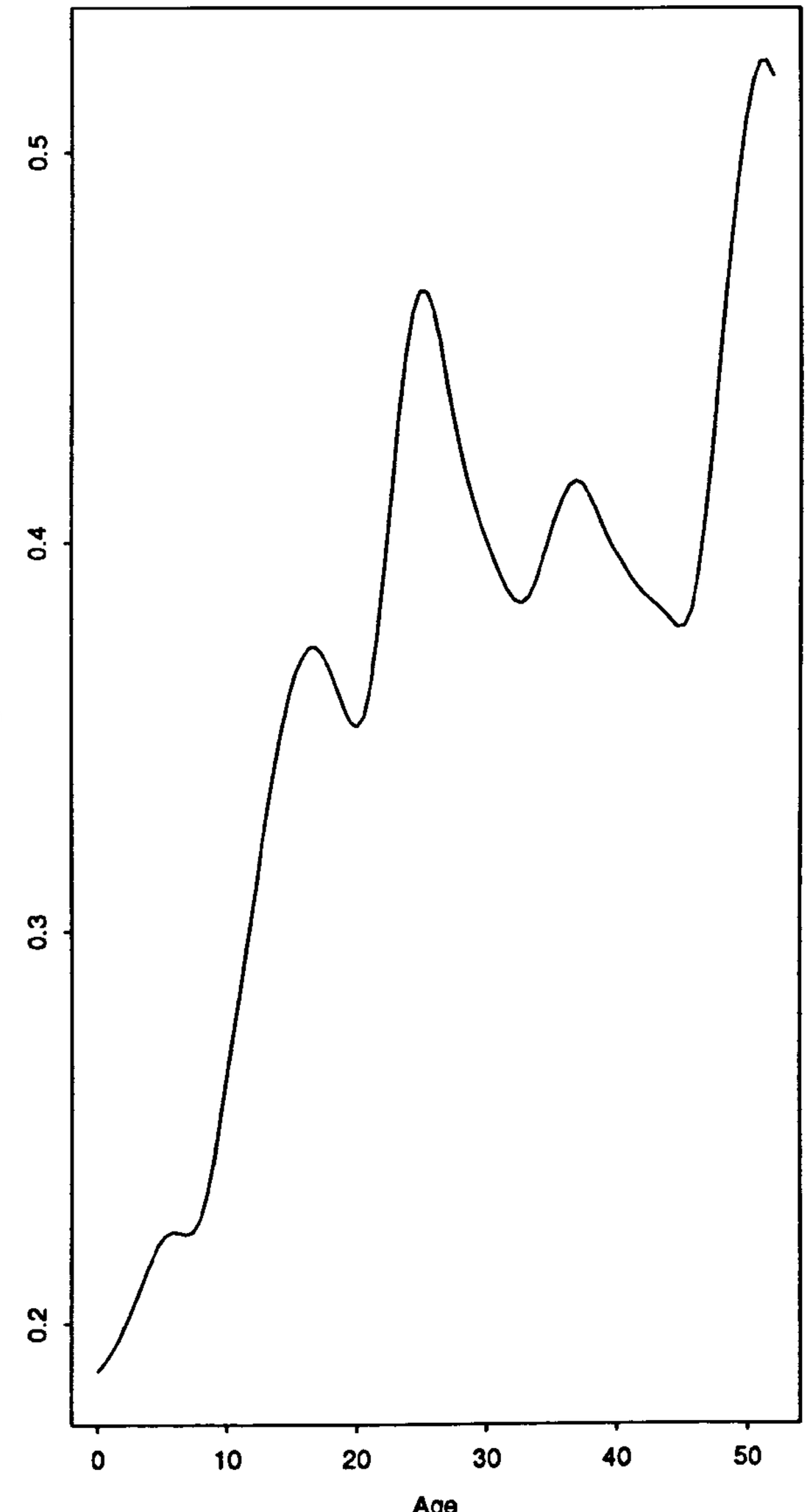
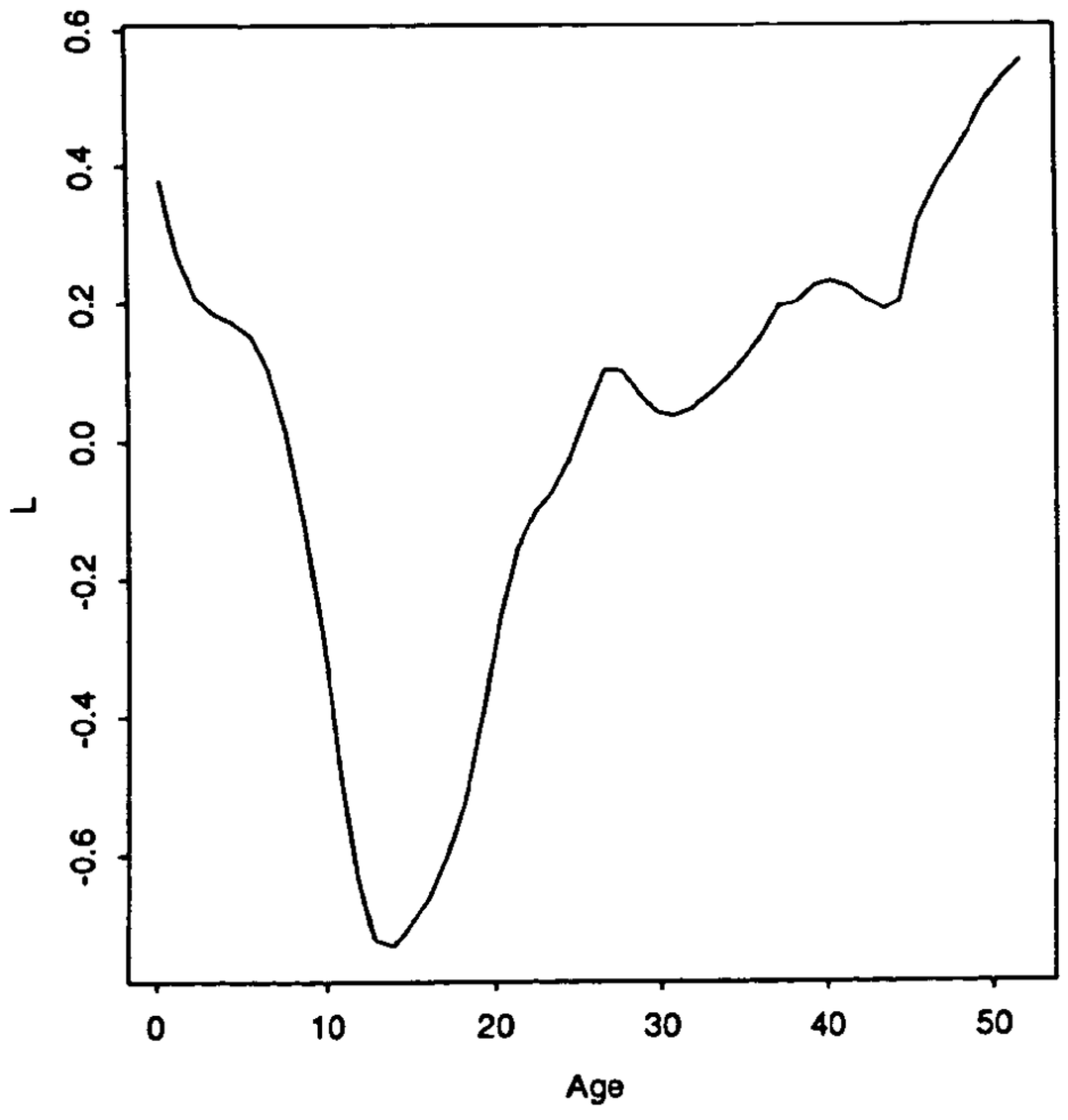
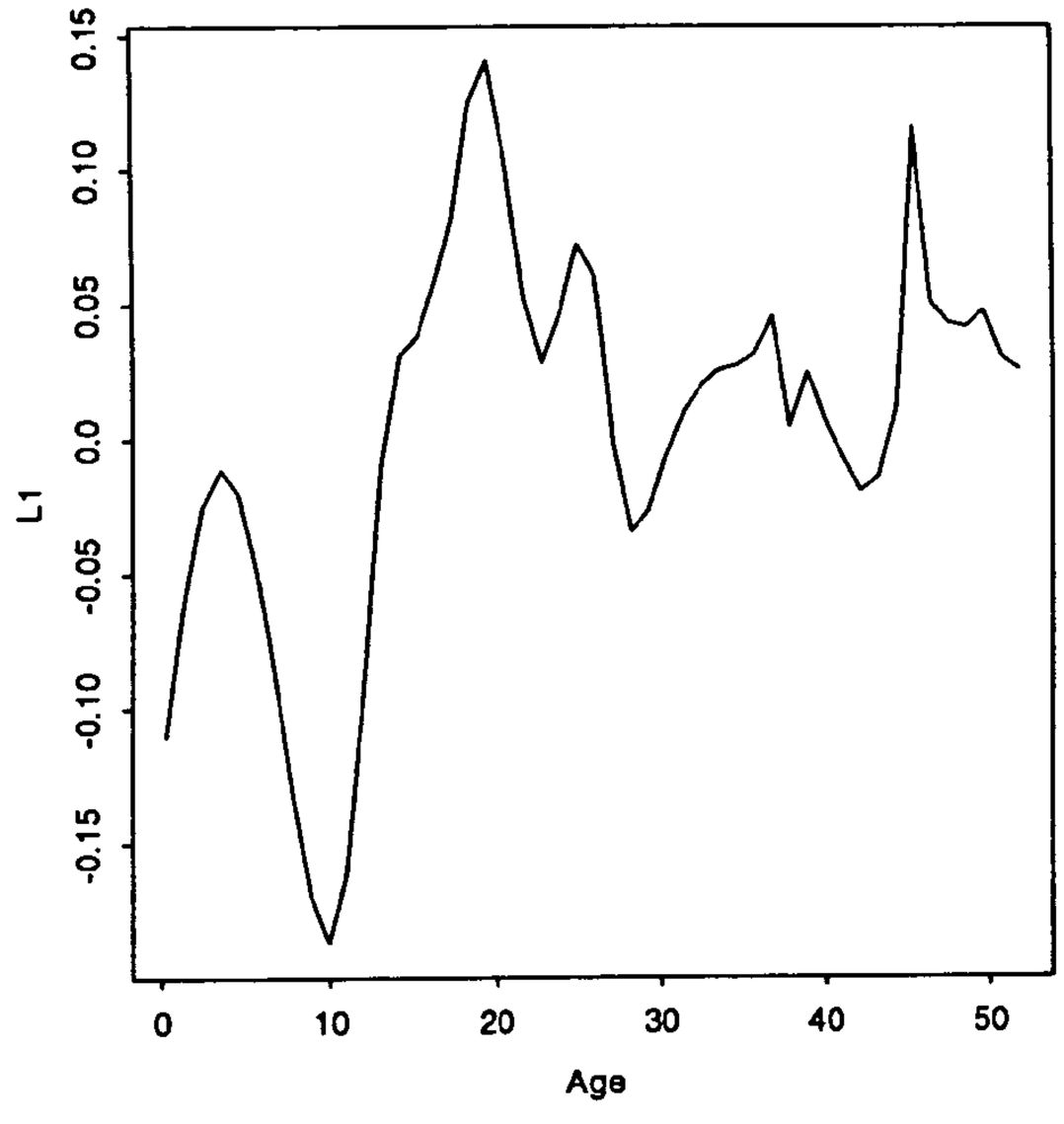


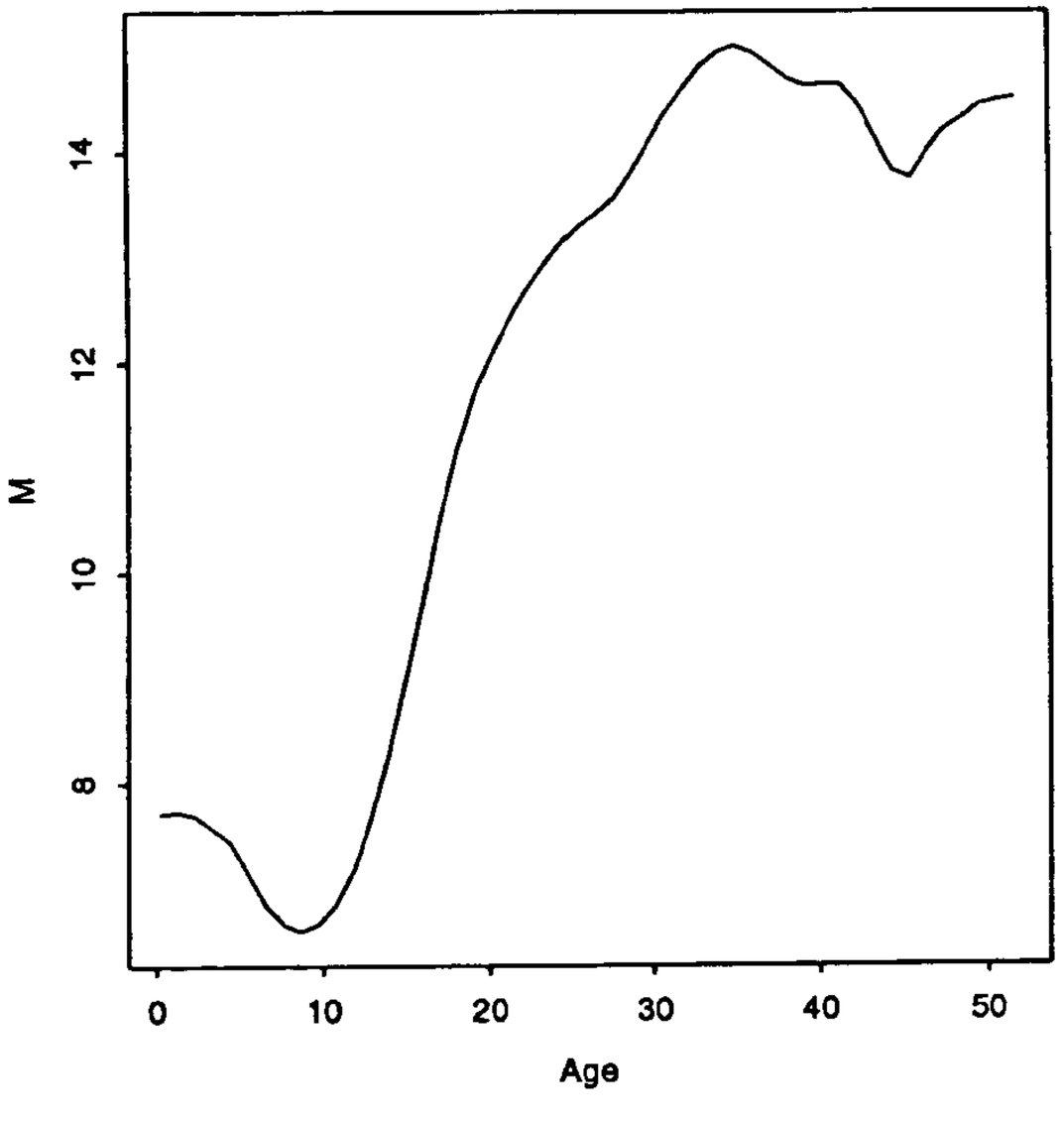
Figure 5.11: Box-Cox power L for Gambian gata by local constant fitting

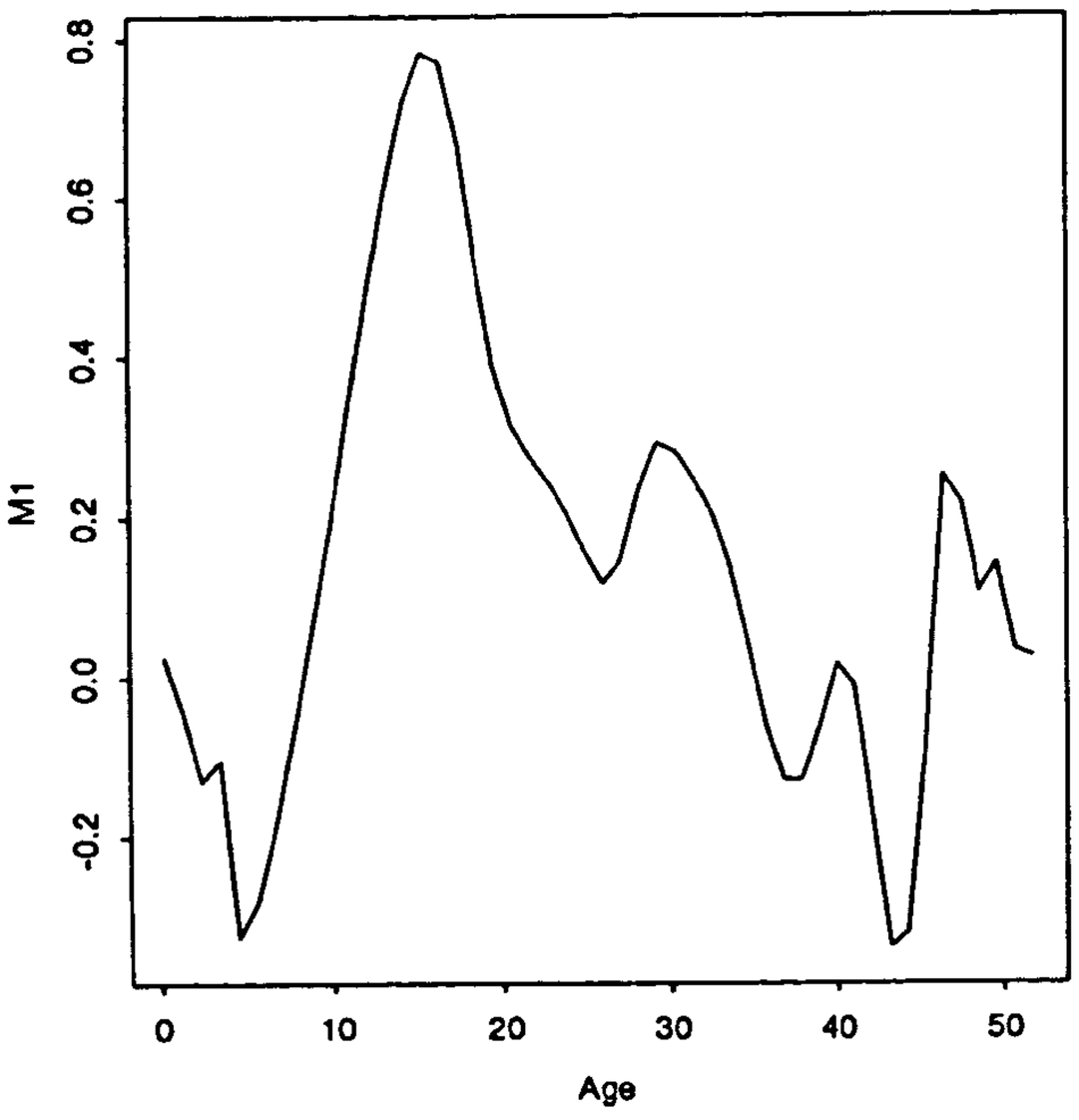


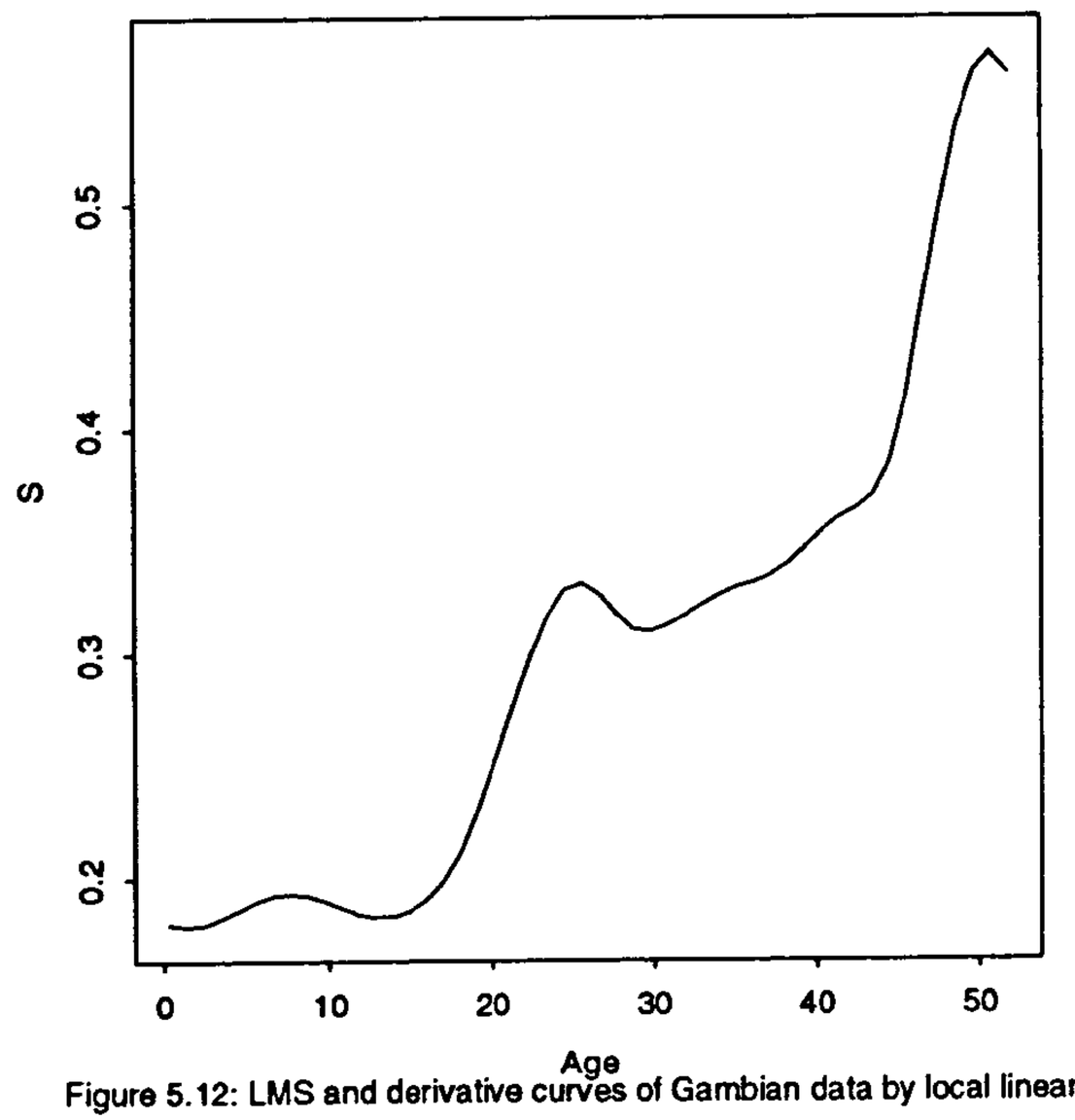












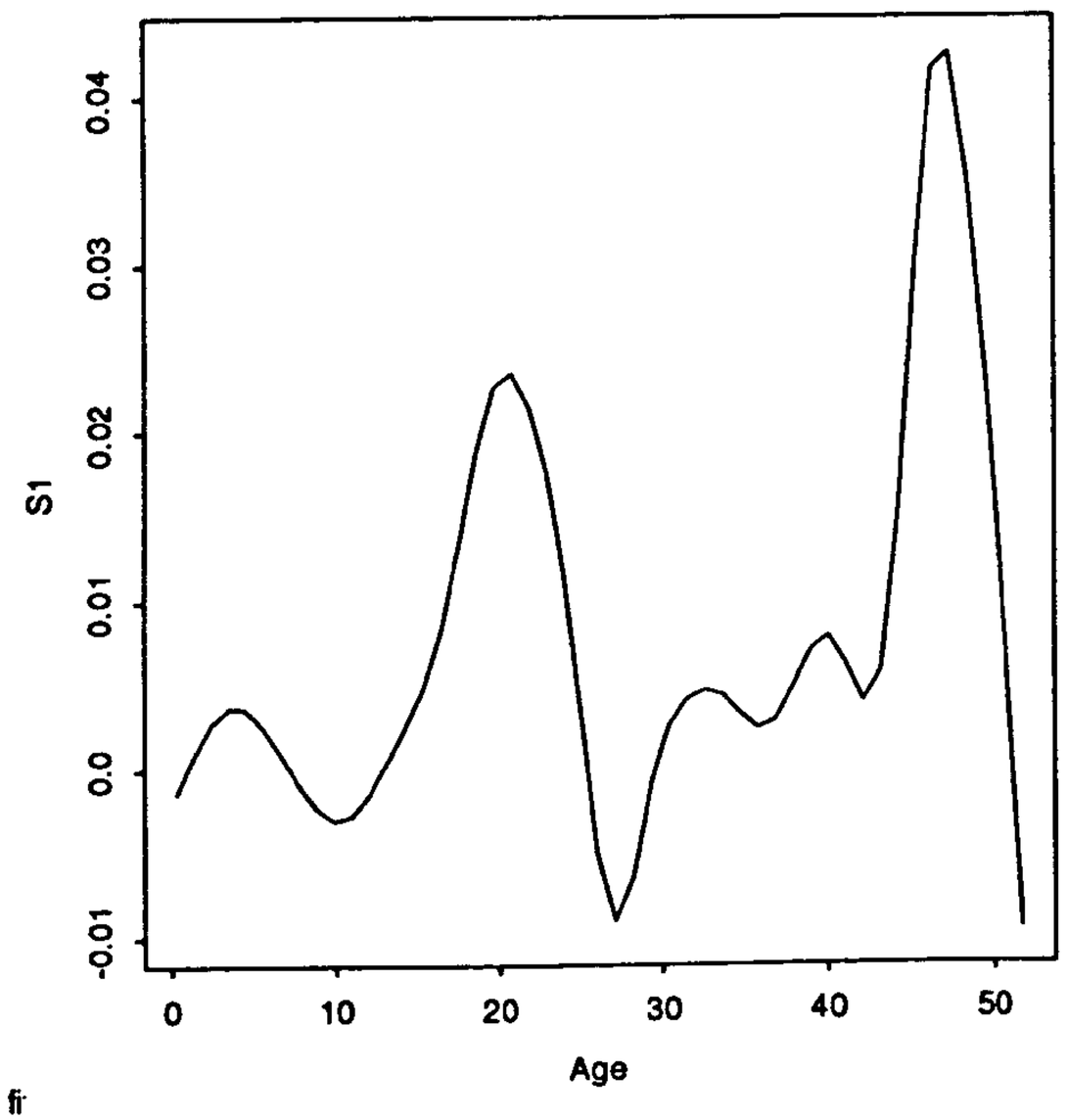


Figure 5.12: LMS and derivative curves of Gambian data by local linear fi

Chapter 6

Quantile Smoothing by Combining NN Estimation with Local Linear Kernel Fitting

6.1 Introduction

Ideally, good quantile regression curves should satisfy some basic requirements, e.g. smoothing with respect to the covariate, goodness-of-fit, concise asymptotic mean squares and convenient computation as possible. A method that satisfies some of these requirements introduced by Bhattacharya and Gangopadhyay (1990) known as Nearest-Neighbor method is further developed in this chapter.

Let $\{(X_i, Y_i), i = 1, 2, ..., n\}$ be iid as (X, Y), and given $X = x_0$, define $Z = |X - x_0|$. Here $\{(Z_i, Y_i), i = 1, 2, ..., n\}$ are iid from (Z, Y). Further the order statistics of Z are denoted by $Z_{n1} < Z_{n2} < ... < Z_{nn}$ and the induced order

statistics of Y by $Y_{n1}, ..., Y_{nn}$, i.e., $Y_{ni} = Y_j$ if $Z_{ni} = Z_j$.

For any positive integer $k \leq n$, the $k - NN$ estimator $\bar{q}_p(x)$ of the conditional p-quantile $q_p(x)$ of Y given $X = x_0$ is the p-quantile of the empirical distribution of conditionally independent responses $Y_{n1}, ..., Y_{nk}$ with cdf

$$\hat{G}_{nk}(y) = k^{-1} \sum_{i=1}^{k} I(Y_{ni} \leq y),$$

and

$$\bar{q}_p(x_0) = the \ \ [kp]th \ \ order \ \ statistic \ \ of \ \ Y_{n1}, ..., Y_{nk} \qquad (6.1)$$

where $I(S)$ denotes the indicator of the event S.

This $k-NN$ estimator has nice Bahadur-type expression as the ordinary quantiles (Bahadur, 1966, Bhattacharya & Gangopadhyay, 1990).

Unfortunately, like in $k-NN$ density estimation (Silverman, 1986), the practical performance of $k - NN$ conditional quantile estimation is not satisfactory by the above criteria. A Monte Carlo example by Healy *et al* (1988) throws enough light on this problem as follows.

The data $\{(X_i, Y_i)\}_1^n, n = 500$ are simulated from the model

$$Y_i = X_i^2 + 10\epsilon_i, \quad \epsilon \sim N(0,1), \quad X_i \sim U(0,10).$$

The median curve is estimated (Figure 6.1) based on $k-NN$ method. Obviously, it is prone to local noise for small k, while it has a heavy tail (right boundary here) for larger k. We found that repeated simulations gave the same result as Figure 6.1.

However, a technique of smoothing is to select not too large k which provides an "initial estimator" of true quantile even it is either prone to large variance or

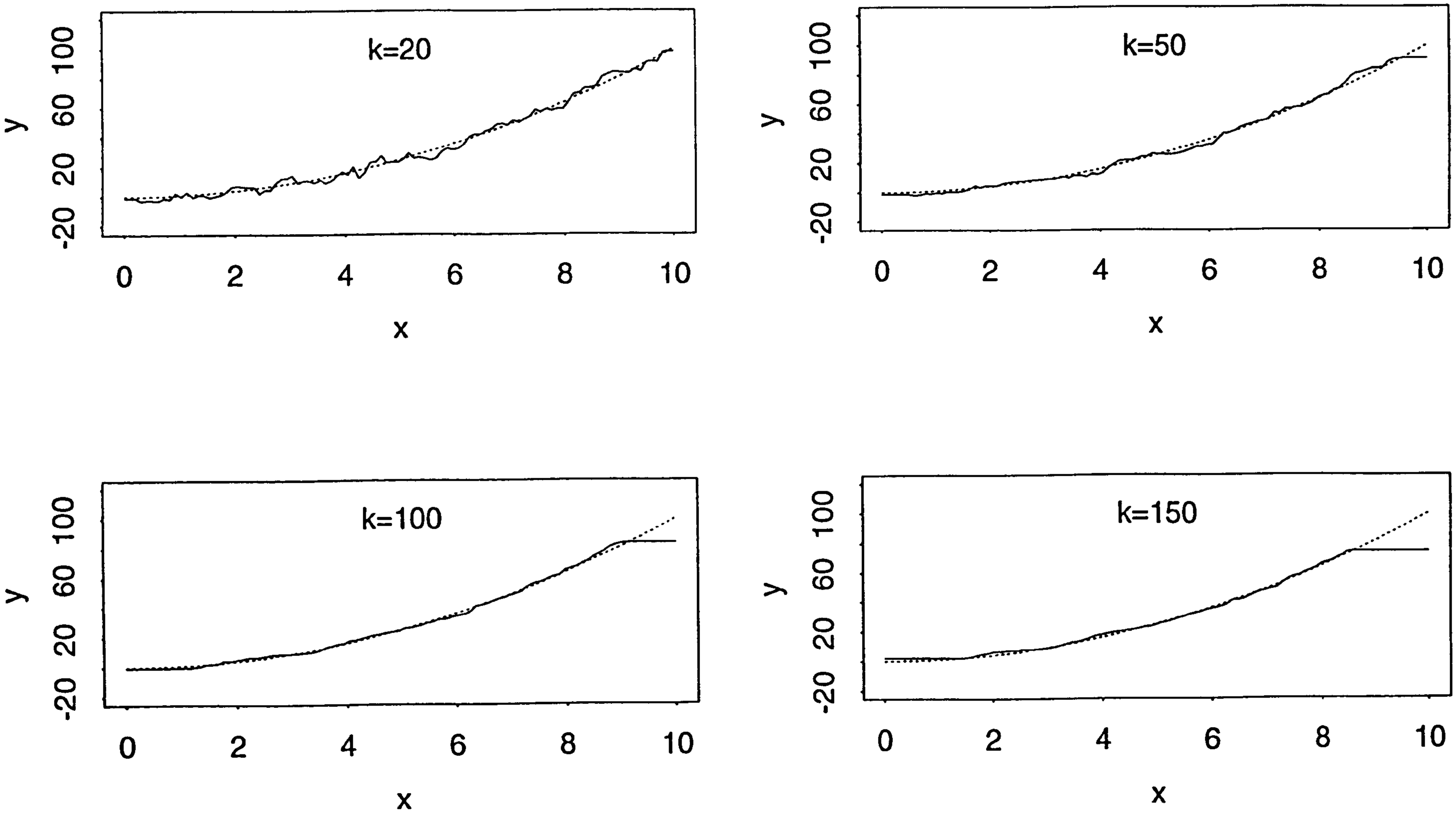


Figure 6.1: Median curve for model $y=x^2+e$, $e \sim N(0,10^2)$, n=500, by k-NN method

boundary bias, then combine this initial estimator with the advanced local linear fitting to get the final smooths. That is, taking this initial quantile as new samples and smooth them again by other smoothing techniques. Among other smoothing techniques, the local linear kernel fitting without boundary modification is a good choice.

On the other hand, it is clear that quality of the fitted curves largely depends on the choice of initial samples regardless of smoothing technique applied and smoothing the $k-NN$ points directly seldom give good goodness-to-fit in terms of the quality of $k-NN$ estimators. The rule provided by Healy *et al* (1988) can be a good help in this context, since this rule successively and repeatedly take the advantage of original sample information when partitioning the covariate-range into boxes. We will call it as HRY partition rule and we will see the rule is different to the general age-grouping and binning data in usual statistical analysis.

In short, a method is developed for quantile fitting that involves two-step smoothing, conveniently called box partition kernel method (BPK).

First produce a set of initial quantiles by $k-NN$ at each covariate point. To do this, first, sort the data by X and denote it by $\{(X_i, Y_i)\}_1^n$, and the sorted $\{Y_i\}_1^n$ can be treated as conditionally independent for $X = x$. The $k-NN$ estimator of p-quantile $q_p(x)$, for given p and k and for any $X = x$ are based on measurements $\{(X_{i+j-1}, Y_{i+j-1})\}_1^k$ $(j = 1, 2, ..., n-k+1)$, that is, the HRY rule partitions the covariate space into $n-k+1$ boxes where the first k points yield the initial estimator at $X = x$, then the procedure is repeated using points 2 to $(k+1)$, 3 to $(k+2)$, ... until the entire covariate space has been covered .

Note that the sample quantiles arising from this first step are irregular and are highly correlated with each other.

Secondly, a smoothing technique is applied using a local linear kernel fitting with properties such as good boundary performance and flexible distribution assumption and selected on the basis of sample quantiles among the covariate values and their $k-NN$ estimators. The median of each box and its $k-NN$ estimators are calculated then used as new sample (sometimes, when the medians of the first box and last box are a bit of far from the extreme points of X, we just take the extreme points as the first and last X observations of new sample, so as to make sure the smooths cover the whole range of X). In doing so, only $n-k+1$ initial sample quantiles are available for smoothing in this step.

Compare to HRY method, this method is much more flexible and has theoretic and computational advantages. Section 6.2 will explore these in details. Section 6.3 will concentrate on asymptotic theory by developing local linear kernel smoothing regression for correlated errors, and we will see that the asymptotic MSE of this method is very concise. Section 6.4 will address the computation and practical performance, and the results are satisfactory.

6.2 Comparison of BPK Method with Other Related Methods

In the first step of HRY method, the data are sorted in ascending order of covariate values, and a set of $(n-k+1)$ boxes each of size k datapoints is made. The minimum box-size $k = max\{\frac{0.5}{p_l}, \frac{0.5}{(1-p_u)}\}$, where p_l and p_u are the lowest and the highest quantiles to be estimated. A regression model is fitted to the first k points in the data set then the required quantiles are obtained from the ranked residuals and are plotted against the median of the k points of the covariate value. This

procedure uses 1 to k points of the data and is repeated using points 2 to $(k+1)$, 3 to $(k+2)$... until the entire covariate space is covered.

After the first step, the sequence $\{X_i^p, Y_i^p\}_1^m$ $(m = n-k+1)$ defines a new set of observations for a p-quantile of the response Y and the corresponding covariate X. The new responses and those $\{Y_i^p\}_1^m$ are neither irregular nor smooth and it is reasonable to assume that they follow a regression model with true function $q_p(x)$

$$Y^p = q_p(X^p) + \epsilon \tag{6.2}$$

Where $E\{\epsilon\} = 0$, obviously, $\{\epsilon_i\}(i \geq 1)$ are correlated errors .

To estimate $q_p(x)$, HRY method may be used in conjunction with polynomial smoothing of the estimate of $q_p(x)$ in each group. To do this, suppose that $q_p(X)$ satisfies

$$q_p(x) = a_{0p} + a_{1p}x + ...a_{tp}x^t.$$

where the coefficients $a_{jp}(j = 0, ..., t)$ are determined by a polynomial function of z_p using least squares fitting, and z_p is the normal equivalent deviate of pth centile, i.e.

$$a_{jp} = b_{j0} + b_{j1}z_p + b_{j2}z_p^2 + ... + b_{jq}z_p^q.$$

Clearly this method is distribution-free, however, it is inadequate in practice as there is tendency of poor fitting at joining points between the groups. To overcome this limitation, Pan, Goldstein and Yang (1990) suggested introduction of extra terms into the smoothing polynomial smoothing of the form

$$q_p(x) = a_{0p} + a_{1p}x + ...a_{tp}x^t + a_{t+1,p}(x - c_1)_+^t + ... + a_{t+l-1,p}(t - c_{l-1})_+^t$$

where c_l is the lth joining point between the groups. Furthermore, Goldstein and Pan (1992) extended the method by simply constructing two or more polynomials joined in a smooth fashion and using generalised least squares estimators, and they showed that the choice of the polynomial order t and q are important and rather depend on experience, simply increasing the order of the polynomial does not work well in general and many experiments must be carried out to the approximate optimal order.

The BPK method is a modification of HRY which uses totally different approach in obtaining the initial quantiles. Also, unlike the above methods it is found that the $k's$ selection has no big influence on the smoothing results as long as the k is not too big ($n-k \to \infty$ when $k, n \to \infty$.) In practice, either $k = max\{\frac{0.5}{p_l}, \frac{0.5}{(1-p_u)}\}$, or double the value works well. Also, BPK uses kernel regression mean method to smooth the estimator of $q_p(x)$ in the regression model (6.2), which does not involve selection of polynomial-type and their orders.

The asymptotic mean square error of BPK will show that a ready-made bandwidth selection rule (Chapter 2) works.

6.3 The Theoretic Model and Mean Square Error (MSE)

For a random vector (X, Y), let g denote the pdf of X and $f(.|x)$ the conditional pdf of Y given $X = x$ with corresponding conditional cdf $F(.|x)$, then

$$F(q_p(x)|x) = p.$$

The sample p-quantiles Y_i^p $(i = 1, ..., m)$ as defined in the previous section, are obtained from k conditionally independent samples $\{Y_i, ..., Y_{i+k-1}\}$ which are part of the original sample $Y_1, Y_2, ..., Y_n$ $(i = 1, ..., m$, and $k + m - 1 = n)$.

Then, the conditional empirical cdf of Y is

$$F_{i,k}(y) = 1/k \sum_{j=1}^{k} I(Y_{k,j+i-1} \leq y) \tag{6.3}$$

which is a general form of (6.1). Then Y_i^p is $[kp]$th order statistic of $Y_{k,i}, Y_{k,i+1}, ..., Y_{k,i+k-1}$ which are induced order statistics of $(Z_i, Y_i), ..., (Z_{i+k-1}, Y_{i+k-1})$, and

$$Y_i^p = inf\{y : F_{i,k}(y) \geq [kp]/k\} \quad i = 1, 2, ..., m.$$

Obviously, for $i = 1$, the statistics Y_1^p and Y_{1+j}^p for $j = 1, 2, ..., k-1$ are related to $\{Y_l\}_{l=j+1}^{k}$, and for any $u > k$, Y_1^p and Y_u^p are independent.

On the other hand, from the Theorem N1 of Bhattacharya and Gangopadhyay (1990), $Y_i^p (1 \leq i \leq m)$ has a Bahadur-type representation as a sum of k independent error random variables. This is

$$Y_i^p - q_p(x) = \beta(q_p(x)) + \frac{1}{kf(q_p(x)|x)} \sum_{j=i}^{i+k-1} W_j(x) + R_k \tag{6.4}$$

where

$$\beta(q_p(x)) = -\frac{g(x)F^{2,0}(q_p(x)|x) + 2g'(x)F^{1,0}(q_p(x)|x)}{24g^3(x)f(q_p(x))}$$

and asymptotically

$$max_{k \in N}|R_k| = O(k^{-3/5}logk),$$

and for each k, the $\{W_j(x_0)\}_i^{i+k-1}$ are independent random variables with mean 0 and variance $p(1-p)$.

Now define

$$T_i = \frac{1}{kf(q_p(x)|x)} \sum_{j=i}^{i+k-1} W_j(x) \tag{6.5}$$

Then clearly

$$var(T_i) = \frac{p(1-p)}{kf^2(q_p(x)|x)} \tag{6.6}$$

and for any $\mu = 0, 1, ...$ the covariance of T_i and $T_{i+\mu}$ depends only on μ. In fact, when $i = 1$,

$$Cov(T_1, T_{1+\mu}) = \begin{cases} \frac{(k-\mu)p(1-p)}{k^2 f^2(q_p(x)|x)} & \mu = 0, 1, ..., k-1 \\ 0 & \mu \geq k \end{cases}$$

Then for any k and sufficiently large n

$$\begin{aligned} \sum_{\mu=1}^{\infty} Cov(T_1, T_{1+\mu}) &= \sum_{\mu=1}^{k-1} Cov(T_1, T_{1+\mu}) \\ &= \frac{(k-1)p(1-p)}{2kf^2(q_p(x)|x)} \end{aligned}$$

This completes the proof of Theorem 6.1 below :

Theorem 6.1: A new sample $\{(X_i^p, Y_i^p)\}_1^m$ is generated from a random sample of n ordered pairs $\{(X_i, Y_i)\}_1^n$ by HRY partition rule and $k-NN$ method for fitting a quantile function $q_p(x)$ in model (6.2), then the errors constitute a stationary process with covariate function

$$E\{\epsilon_{n,i}, \epsilon_{n,j}\} = \sigma^2(x)\rho_n(|i-j|) \tag{6.7}$$

where $\sigma^2(x)$ is given by $\frac{p(1-p)}{kg^2(q_p(x)|x)}$ and

$$\rho_\mu = \begin{cases} \frac{k-\mu}{k} & \mu = 1, 2, ..., k \\ 0 & \mu \geq k \end{cases}$$

Theorem 6.2. Given n pairs of iid observations $\{(X_i, Y_i)\}_1^n$, and under the conditions of Theorem 6.1 and if $h \to 0$, $nh \to \infty$, then the local linear kernel estimator $\hat{q}_p(x)$ of $q_p(x)$ based on regression model (6.2) with 2nd order symmetric kernel satisfies

(i) Interior property:

$$E(\hat{q}_p(x) - q_p(x))^2 \approx 1/4(q_p''(x))^2\mu_2^2(K)h^4 + \frac{R(K)}{mh}(\frac{p(1-p)}{kf(q_p(x)|x)^2}(1+2\sum_{\nu}\rho_\nu))$$
$$= 1/4(q_p''(x))^2\mu_2^2(K)h^4 + \frac{R(K)p(1-p)}{mhf(q_p(x)|x)^2}.$$

(ii) Boundary behavior: Assume $x \in [0,1]$, then for left-boundary points $x = ch$ with $c > 0$,

$$E(\hat{q}_p(x) - q_p(x))^2 \approx 1/4(q_p''(0+))^2\{\frac{s_{2,c}^2 - s_{1,c}s_{3,c}}{s_{2,c}s_{0,c} - s_{1,c}^2}\}^2h^4$$
$$+ \frac{\int_{-\infty}^{c}[s_{2,c} - us_{1,c}]^2K^2(u)du}{[s_{2,c}s_{0,c} - s_{1,c}^2]^2}(\frac{p(1-p)}{kf(q_p(0+)|0+)^2}(1+2\sum_{\nu}\rho_\nu))$$
$$= 1/4(q_p''(0+))^2\{\frac{s_{2,c}^2 - s_{1,c}s_{3,c}}{s_{2,c}s_{0,c} - s_{1,c}^2}\}^2h^4h^4$$
$$+ \frac{\int_{-\infty}^{c}[s_{2,c} - us_{1,c}]^2K^2(u)du}{[s_{2,c}s_{0,c} - s_{1,c}^2]^2}\frac{p(1-p)}{mhf(q_p(0+)|0+)^2},$$

where $s_{l,c} = \int_{-\infty}^{c} K(u)u^l du, l = 0,1,2,3, m = n - k + 1.$

To prove this theorem the following lemma is need which follows Lemma 2 of Fan and Gijbels (1995).

Lemma. Assume that $g(.)$, $K(.)$ and $S(.)$ are bounded and continuous functions in [0,1] and right continuous at $x = 0$. Further suppose that $\limsup_{u\to-\infty}|K(u)u^{l+2}| < \infty$ for a nonnegative integer l. Then

(i) For interior points $x \in [0,1]$,

$$\sum_{j=1}^{n} K(\frac{x - X_j}{h})S(X_j)(x - X_j)^l = nh^{l+1}S(x)g(x)\int_{-\infty}^{+\infty} K(u)u^l du(1 + 0_P(1)).$$

(ii) For boundary point where $c = xh_n$ (if $c > 0$ is left boundary), when $h_n \to 0$ $(n \to \infty)$,

$$\sum_{j=1}^{n} K(\frac{x - X_j}{h})S(X_j)(x_n - X_j)^l = nh^{l+1}S(0+)g(0+)\int_{-\infty}^{c} K(u)u^l du(1 + 0_P(1)).$$

Proof of Theorem 6.2: Note that

$$\hat{q}_p(x) = \sum_{j=1}^{m} w_{h,m}(x,j) Y_j^p \tag{6.8}$$

where the weights are local linear kernel fitting weights.

$$w_{h,m}(x,j) = \frac{K(\frac{x-X_j^p}{h})(S_{m,2} - (x - X_j^p)S_{m,1})}{S_{m,2}S_{m,0} - S_{m,1}^2},$$

with

$$S_{m,l} = \sum_{1}^{m} K(\frac{x - X_j}{h})(x - X_j)^l, \quad l = 0,1,2,3.$$

Conditioning on covariates X_j^p, $j = 1, ..., m$, and letting $\Sigma_m(.)$ be the covariance of the observations, $\nu(.)$ the column vector $\nu(j)$, and $\nu^T(.)$ the transpose of $\nu(.)$, then

$$MSE(x,h,m,p) = \Big(w_{h,m}(x,.)^T q_p(.) - q_p(x)\Big)^2 + w_{h,m}(x,.)^T \Sigma_m(.) w_{h,m}(x,.).$$

Since the bias term $\Big(w_{h,m}(x,.)^T q_p(.) - q_p(x)\Big)$ is not affected by the correlation structure and has the same asymptotic form as the bias provided by Fan (1992, 1993), thus

(i) for interior point

$$w_{h,m}(x,.)^T q_p(.) - q_p(x) = -1/2 h^2 \mu_2(K) q_p''(x) + o(h^2) + o(1/mh)$$

(ii) for boundary point

$$w_{h,m}(0+,.)^T q_p(.) - q_p(0+) = -1/2 h^2 \alpha(K,c) q_p''(+0) + o(h^2) + o(1/mh),$$

with $\alpha(K,c) = \frac{s_{2,c}^2 - s_{1,c}s_{3,c}}{s_{2,c}s_{0,c} - s_{1,c}^2}$.

$S_{m,l} = mh^{l+1} g(x) s_l (1 + o_P(1))$ at the interior points and

$S_{m,l} = mh^{l+1} g(o+) s_{l,c} (1 + o_P(1))$ at the boundary points.

Since the correlation function $\rho(.)$ is independent of the covariate, and $\sum_j^m |\rho_\mu|$ coverages as $m \to \infty$, and let $S(.) = \sigma^2(.)(1 + 2\sum_{\mu=1}^{\infty} \rho_\mu)$. To derive the variance item consider

$$|w_{h,m}(x,.)^T \Sigma_m(.) w_{h,m}(x,.) - \tfrac{R(K)}{nh}\sigma^2(.)(1+2\rho_\mu)|$$

$$\leq |w_{h,m}(x,.)^T \Sigma_m(.) w_{h,m}(x,.) - w_{h,m}(x,.)^T \sigma^2(.)(1+2\rho_\mu) w_{h,m}(x,.)|$$

$$+|w_{h,m}(x,.)^T \sigma^2(.)(1+2\rho_\mu) w_{h,m}(x,.) - \tfrac{R(K)}{nh}\sigma^2(.)(1+2\rho_\mu)|$$

$$\leq o(1/mh) + \sigma^2(.)(1+2\rho_\mu)|w_{h,m}(x,.)^T w_{h,m}(x,.) - \tfrac{R(K)}{nh}|$$

Then

$$|w_{h,m}(x,.)^T w_{h,m}(x,.) - \frac{R(K)}{nh}| = o(1/mh),$$

and at the interior point

$$MSE(x,h,m,p) = 1/4h^4\mu_2(K)^2 q_p''(x)^2 + \frac{R(K)}{nh}\sigma^2(1+2\rho_\mu) + o(1/nh) + o(h^4) \quad (6.9)$$

and hence (i).

The (ii) in boundary points $x = ch$ with $c > 0$ and $h \to 0$ can be proved along same lines as (i) .

Remark 1. From the asymptotic pointwise mean square error (MSE), it is seen that BPK smooth gives the same results as direct minimization of "check function" by local linear kernel fitting (Fan, Hu and Truong, 1994, Jones and Hall, 1990), but BPK is more easily computed.

Remark 2. An interesting feature is that the MSE is independent of k, but asymptotically it is required that as $n \to \infty$, $k \to \infty$, and $n - k \to \infty$.

Remark 3. The conclusions are not limited to fixed design.

6.4 Bandwidth Selection and Numerical Example

The optimal bandwidth for interior points could be obtained from (i) of Theorem 6.2 as

$$h^5 = \frac{R(K)p(1-p)}{f^2(q_p(x)|x)q_p''(x))^2\mu_2^2(K)m} \tag{6.10}$$

As mentioned in *Remark 1* above this method of combining $k-NN$ estimation with local linear kernel mean fitting for smooth conditional p-quantile is asymptotically equivalent to local linear kernel weighting "check-function" based on m independent samples in the sense of MSE. Therefore, the same rule-of-thumb is available:

(a) use ready–made, and sophisticated, methods to select h_{mean} such as the technique suggested by Ruppert, Sheather & Wand (1995);

(b) use $h_p = h_{mean}\left\{\frac{p(1-p)}{\phi(\Phi^{-1}(p))^2}\right\}^{1/5}\left(\frac{n}{m}\right)^{1/5}$ to obtain all other h_ps from h_{mean}.

The choice of k plays an important role in applying the GROSTAT program (Rasbash and Pan, 1990). It is suggested $k \geq 50$ and a minimum box size the GROSTAT program accept to implement HRY procedure is $max\{0.5/p_l, 0.5/(1-p_u)\}$.

However, avoiding strict requirement on k, the minimum limit or its two to four times all seem to work well in practice. The method is applied to fit a set of quantiles using different values of k:

First data are simulated (n=500) from the model of Section 6.1. A normal kernel and $h_{mean} = 1$ selected subjectively are used to fit the median with $k = 10, 20, 50$ and 100. It is seen from Figure 6.2 that $k \geq 50$ is not necessary for BPK method and different k has small smoothing effect on the fitted curve. Figure 6.2 is based on 100 simulations. Further, the method is used to fit $5th$, $10th$, $25th$, $50th$, $75th$, $90th$ and $95th$ percentile for $k = 20$ which is displayed in Figure 6.3.

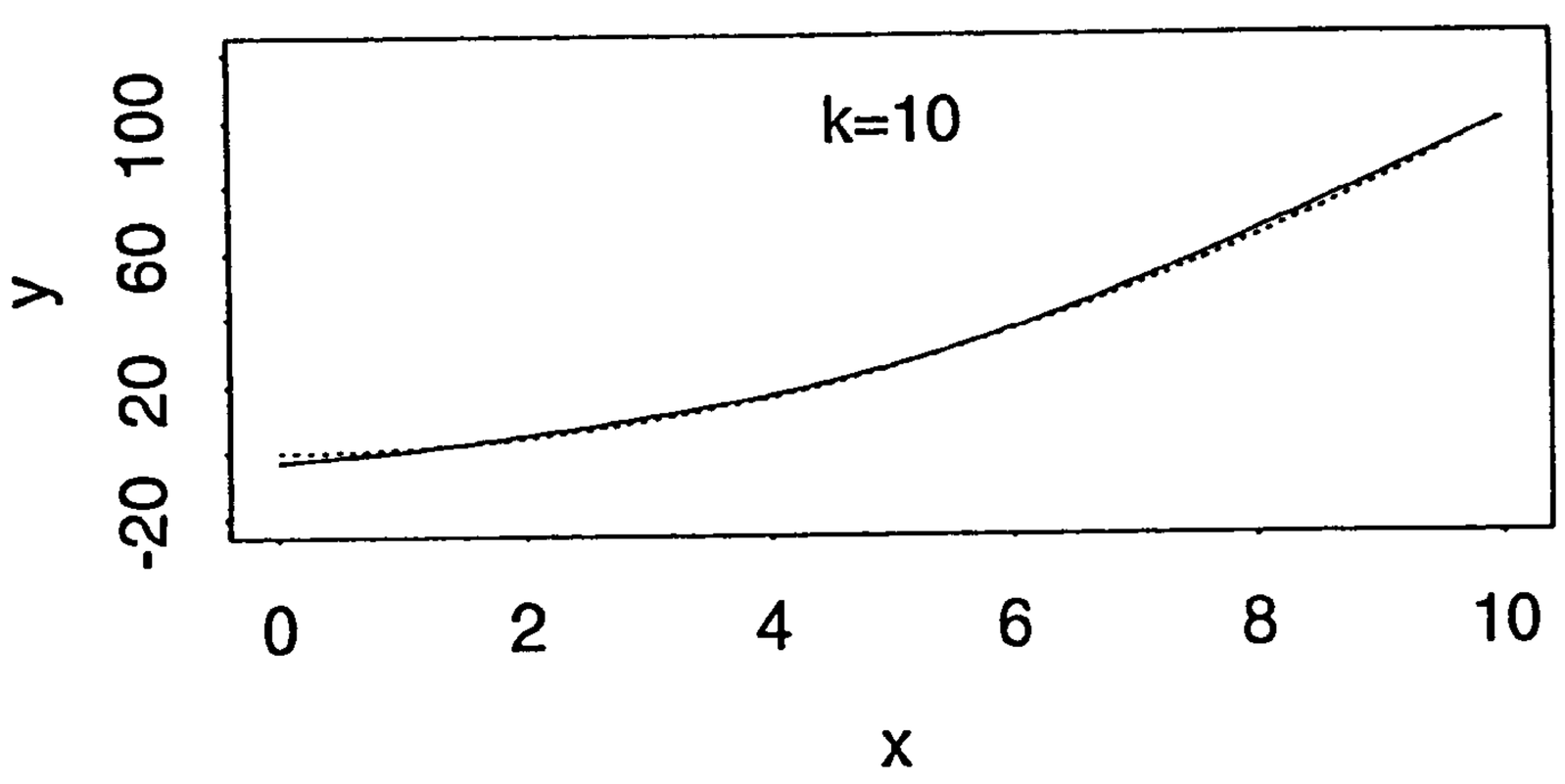
Secondly, for

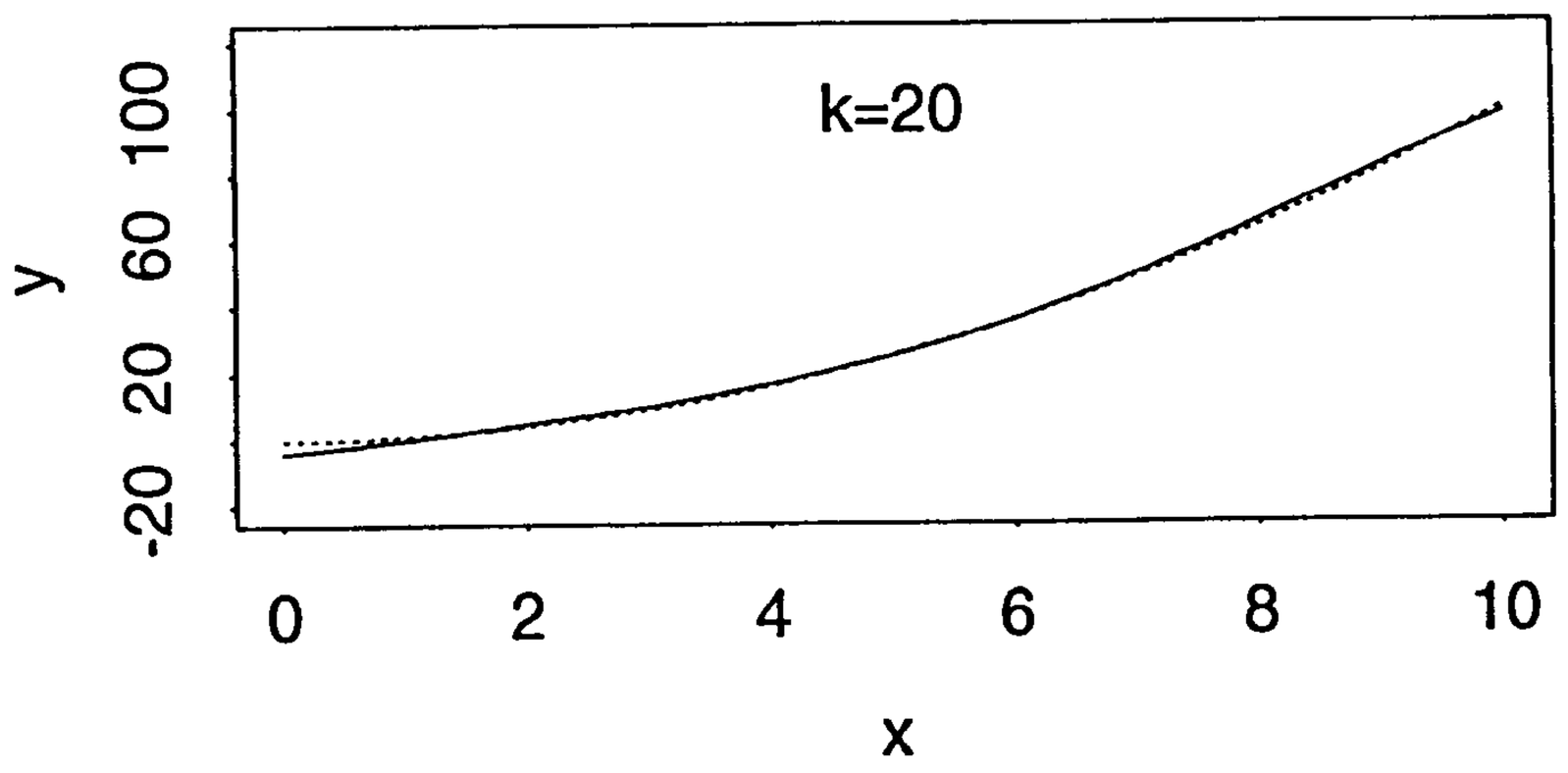
(i) Gambian data with $n = 892$, and for $h_{mean} = 1.09$ and $k = 30, 50$. The seven fitted quantiles are $\{p = 0.5, 0.25, 0.75, 0.9, 0.1, 0.97, 0.03\}$ (Figure 6.4).

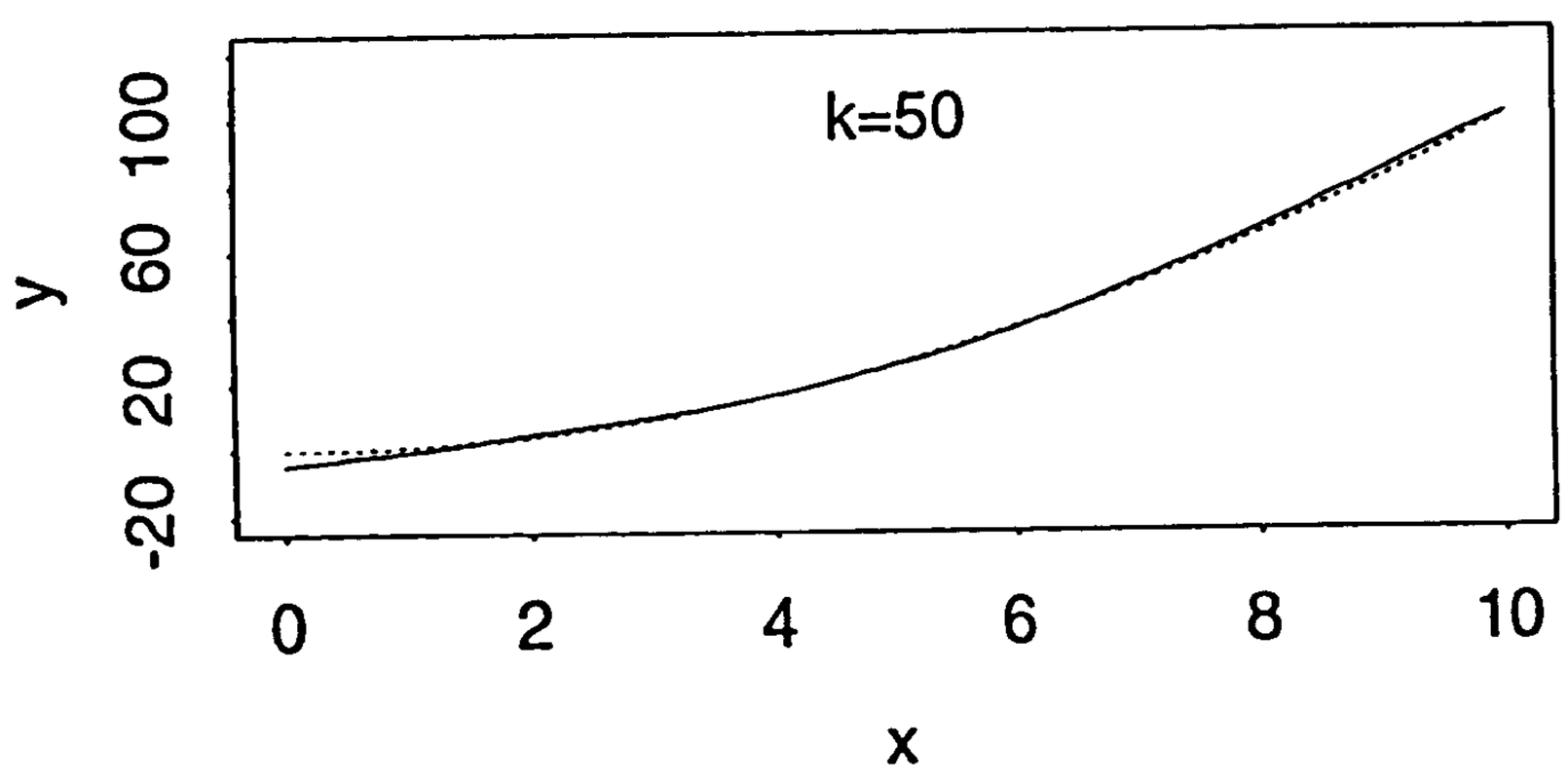
(ii) US girls' weight data: For which $n = 4011$, the same set of quantiles as in (i) is fitted where $h_{mean} = 1.8$ for $k = 30, 50$ (Figure 6.5).

(iii) Serum concentration data (IgG): Here $n = 300$, the quantiles $\{p = 0.5, 0.25, 0.75, 0.9, 0.1, 0.95, 0.05\}$ are fitted where $h_{mean} = 0.5$ and $k = 20, 30$ (Figure 6.6).

For Gambian data, small differences between moderate quantile curves are observed on the right boundary for the two different k values. However, the differences are not very noticeable, and particularly not for sets (ii) and (iii) (see Figures 6.5 and 6.6). Comparing these fits with those by methods in earlier chapters (Chapter 2, 3 and 5), the fitting quality for real data sets is almost as good as double kernel method, better than single kernel and semi-parametric methods.







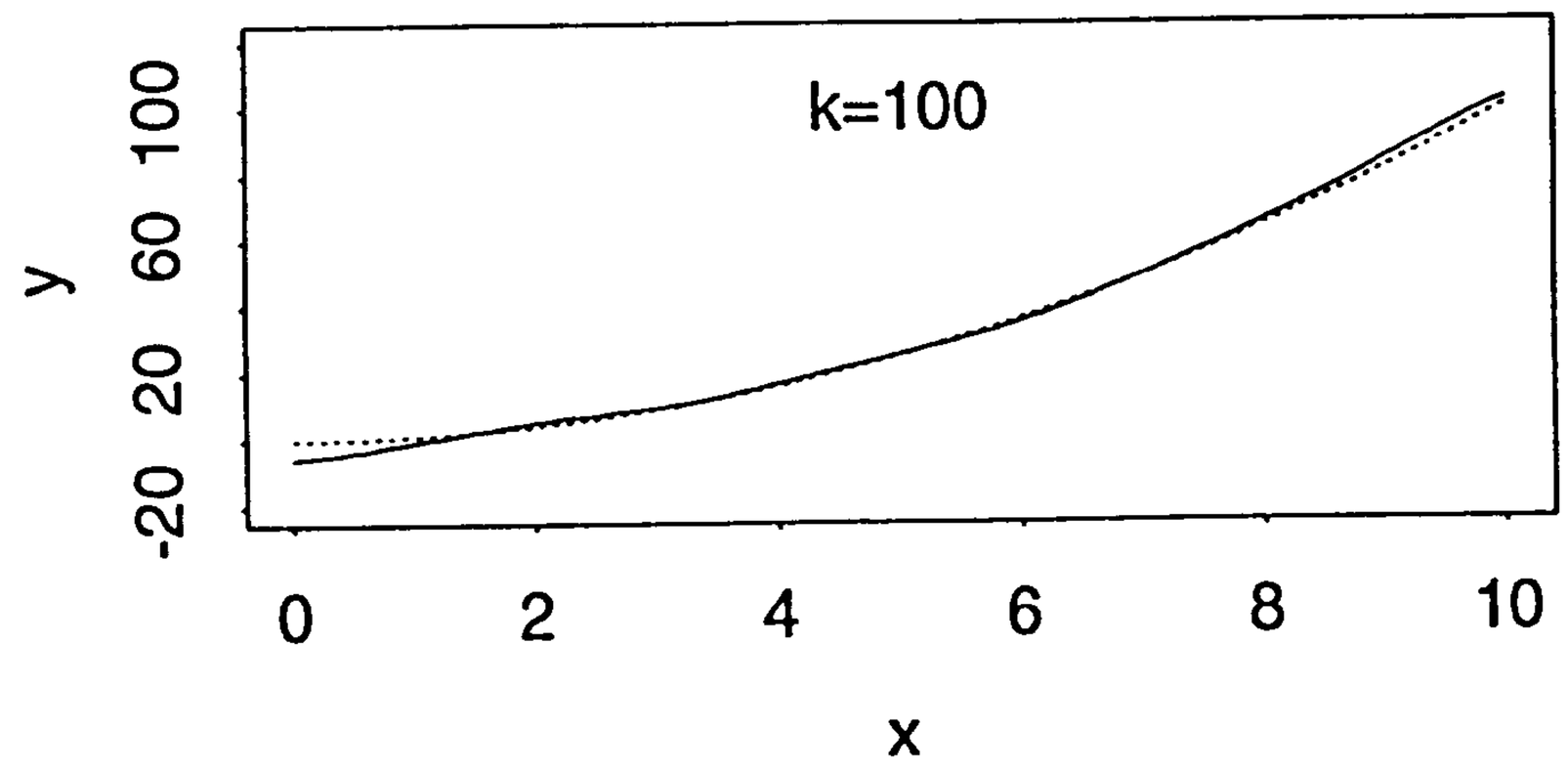


Figure 6.2: Median curve for model $y=x^2+e$, $e \sim N(0,10^2)$ by BPK method with 100 simulations

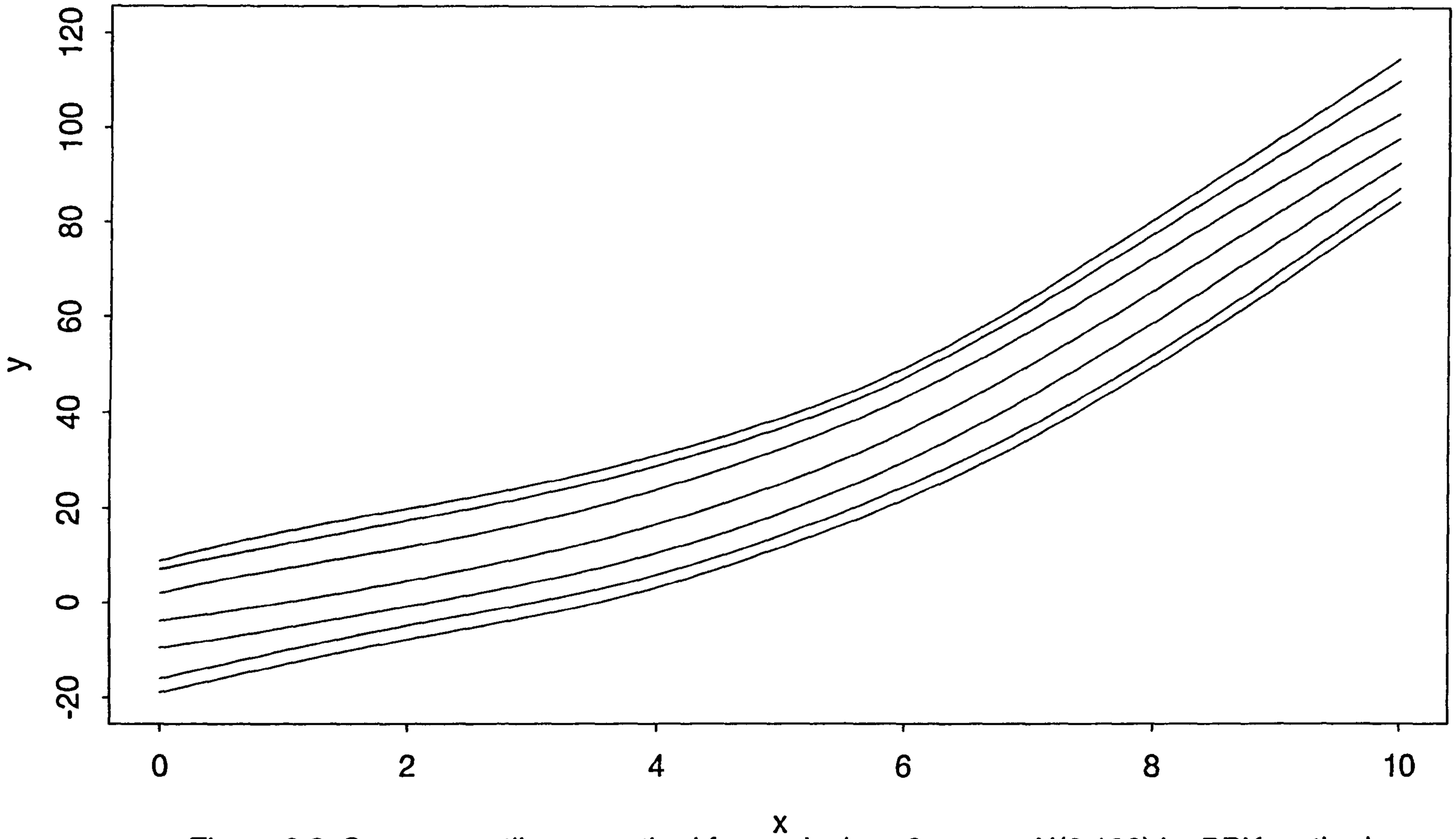


Figure 6.3: Seven quantiles smoothed for modoel y=x2+e , e ~ N(0,100) by BPK method

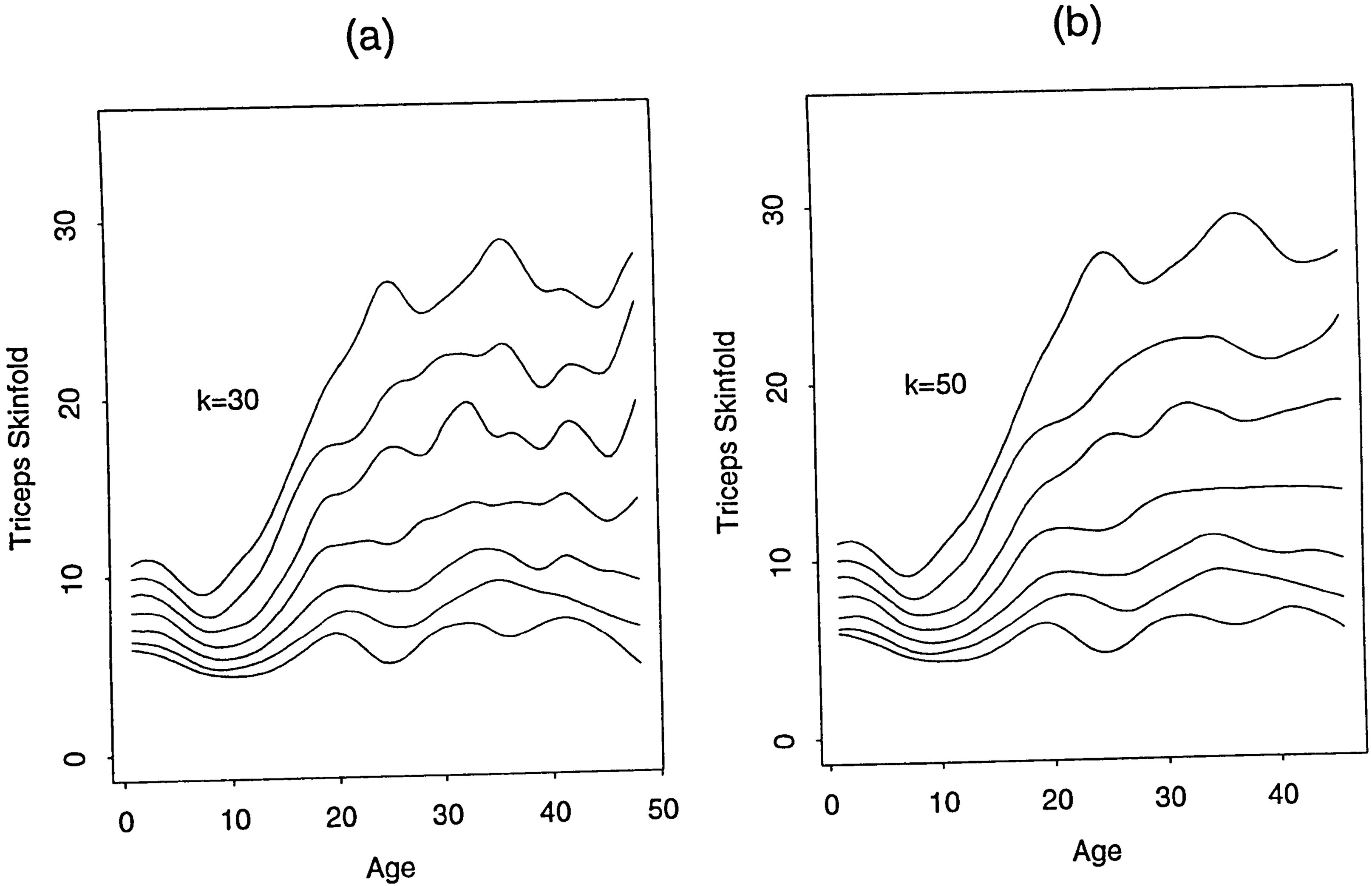


Figure 6.4: Seven quantiles of Gambian data fitted by BPK method

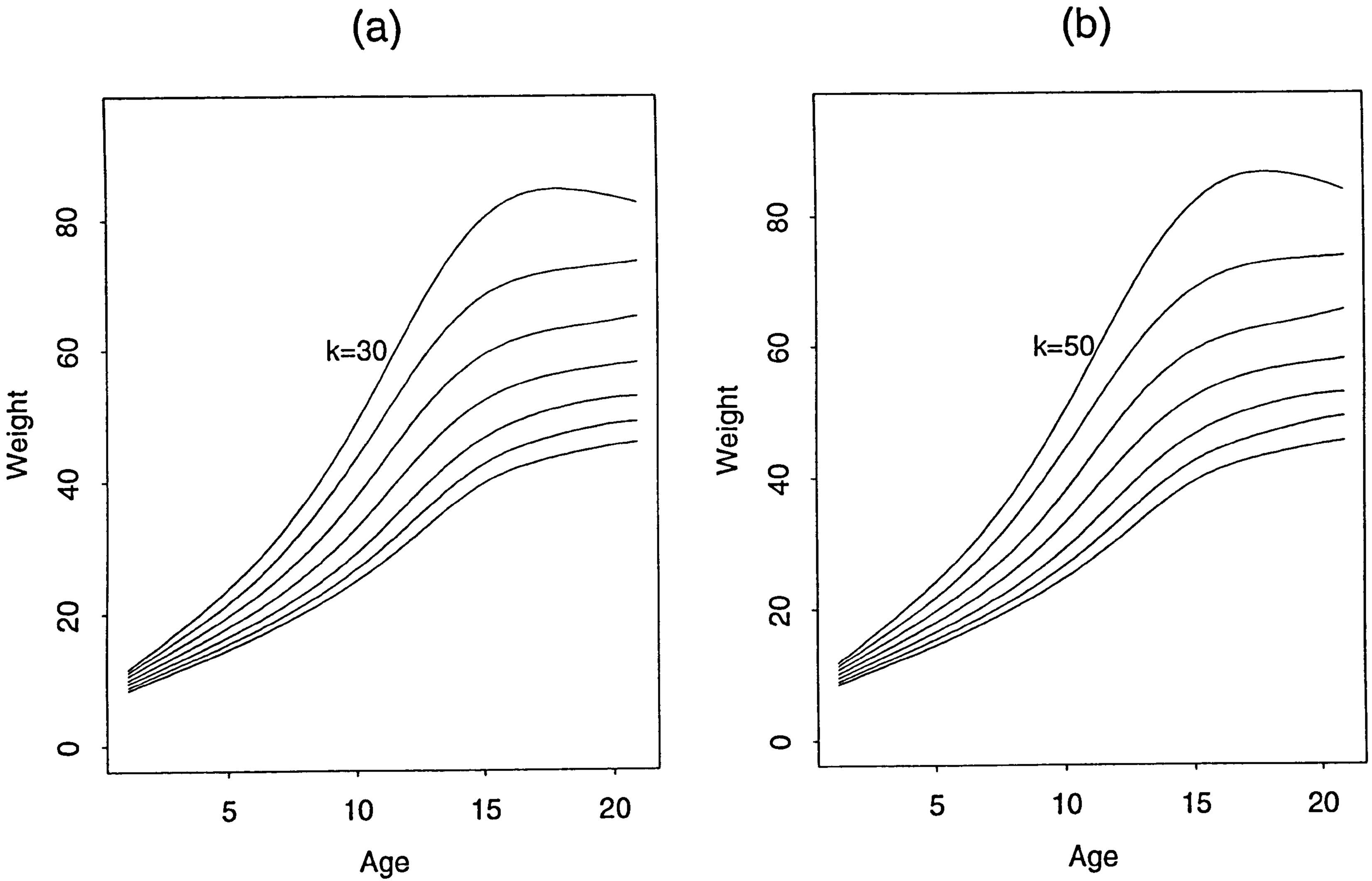


Figure 6.5: Seven quantiles of US data fitted by BPK method

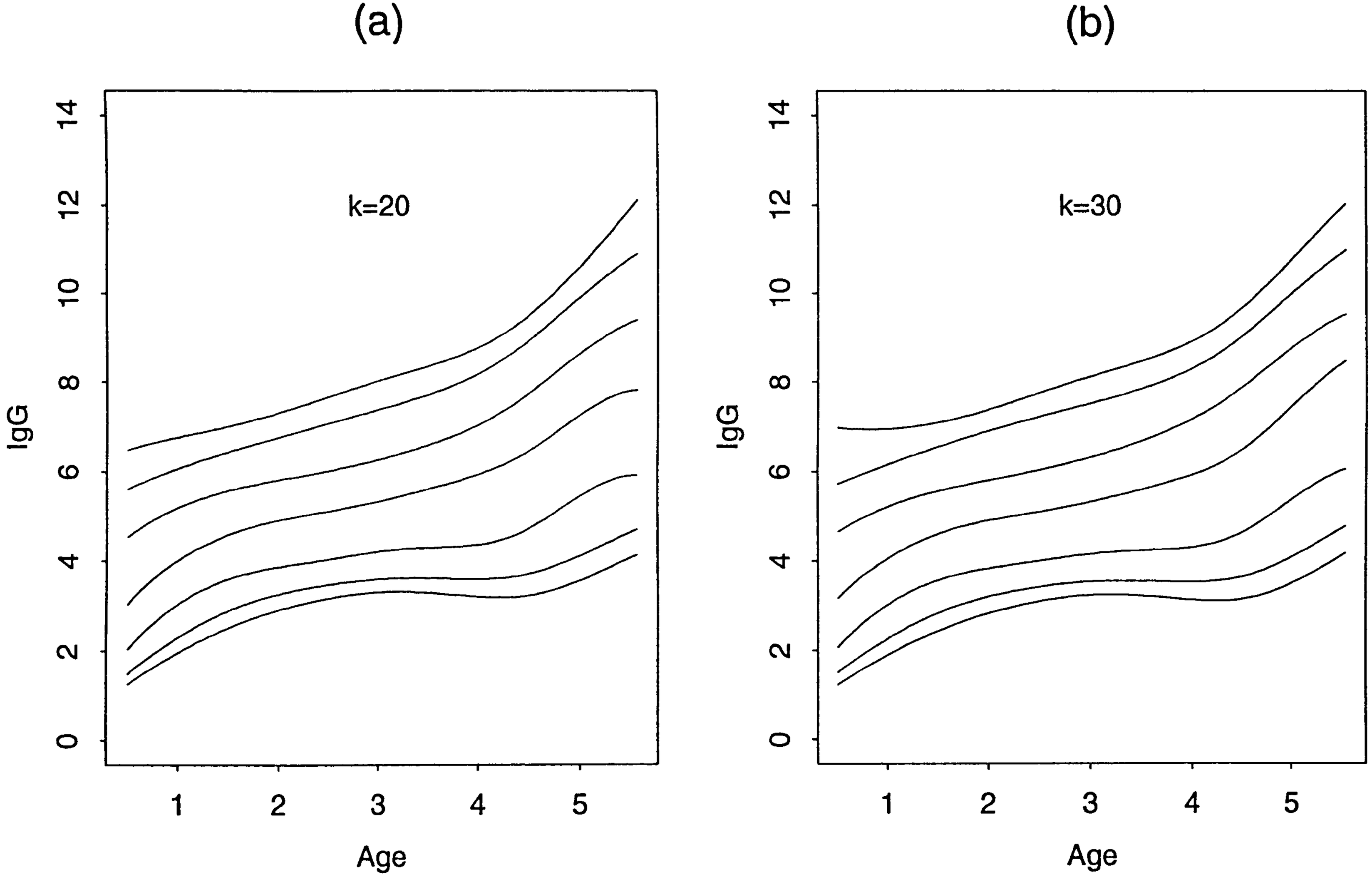


Figure 6.6: Seven quantiles of IgG data fitted by BPK method

Chapter 7

Some Simulation Comparisons of Regression Quantile Methods

7.1 Introduction

As we have seen in previous chapters, smoothing regression quantiles have been studied in a variety of ways under nonparametric or semi-parametric frames, for example, loss-based kernel-weighting, distribution (or density)-based kernel-weighting, likelihood-based kernel-weighting, likelihood-based spline, and nearest-neighbour with some adjustment such as double-kernel instead of single-kernel and local linear fitting instead of local constant fitting. The following list of smoothing methods for regression quantiles as discussed in the relevant chapters cover the whole range.

N0.1. Local constant kernel weighting minimizing "Check function" method (Chapter 3).

N0.2. Local linear kernel weighting minimizing "Check function" method (Chapter 2, Chapter 3).

N0.3. Local constant double kernel method.

N0.4. Local linear double kernel method (Chapter 2, Chapetr 4).

N0.5. LMS method which is the spline-fitting method of Cole and Green.

N0.6. Likelihood-based kernel method (Chapter 5).

N0.7. Two-stage Method or BPK method (Chapter 6).

Estimators $\hat{q}_p(x)$ of the p-quantile $q_p(x)$ are calculated by each of the above seven methods at interior and boundary points of the covariate X using simulated data.

For interior points $a < x < b$ Integrated Square Errors (ISE) of the estimators are used as a measure of closeness which are defined as

$$ISE_p = \int_a^b [\hat{q}_p(x) - q_p(x)]^2 dx$$

where $\hat{q}_p(x)$ is the estimator of pth regression quantile $q_p(x)$ of response Y given $X = x$.

Also the absolute deviation (AD) is used to compare the estimators $\hat{q}_p(x)$ at the boundary points

$$AD_p = |\hat{q}_p(x) - q_p(x)|$$

where $\hat{q}_p(x)$ and AD are calculated at $x = x_L$, x_R, left (L) and right (R) boundaries of x respectively.

In other words, the attention here is shifted from large sample study to small

sample summary. To give a comprehensive comparison, a fundamental frame is given in the following.

Suppose that the covariate and response variables (X, Y) are linked by a general model

$$Y = m(X) + \sigma(X)\epsilon \tag{7.1}$$

where the random error $\epsilon \sim \chi(z)$ is independent of covariate $X \sim g(x)$, and all p-quantiles z_p ($0 < p < 1$) of ϵ exist. Clearly the regression mean and p-quantile $q_p(x)$ are

$$\begin{aligned} E\{Y|X = x\} &= m(x) + \sigma(x)E\{\epsilon\}, \\ q_p(x) &= m(x) + \sigma(x)z_p. \end{aligned} \tag{7.2}$$

When $\sigma(x) = \sigma$ for all x, model (7.1) is called homoscedastic model, otherwise, heteroscedastic one.

Four models with three quantile points ($p = 0.1, 0.5, 0.9$) and two different samples ($n = 100$ and $n = 500$) are simulated, ISE and AD are calculated based on 100 simulations, each of which varies y's and conditional on single set of x's.

Hopefully, in the simulations, a range of versions of the model (7.1), i.e. of $m(x)$, $g(x)$, $\chi(z)$ and z_p is employed which highlight:

1. Linearity of $q_p(x)$ in x;

Note that $q_p(x)$ is a linear function of x when (X, Y) is a bivariate normal and bivariate Γ distribution. (As a matter of fact, when $(X, Y) \sim N(a, b, \sigma_1, \sigma_2, r)$, $q_p(x) = b + \sigma_2\sqrt{1-r^2}\Phi^{-1}(p) + r\frac{\sigma_2}{\sigma_1}(x-a)$, and when (X, Y) has density $f(x, y) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(y-x)^{\beta-1}e^{-y}$, $x > 0, y > x$, $q_p(x) = x + b_p$ with b_p is determined

according to $p\Gamma(\beta) = \int_0^{b_p} z^{\beta-1}e^{-z}dz$.)

2. Quadratic properties of $q_p(x)$:

As seen, asymptotic MSEs of the estimators are largely related to the first and second derivatives of $q_p(x)$.

3. Jumps in $q_p(x)$.

Also, heteroscedasticity of the model, $g(x)$ is low density or high density, $\chi(z)$ is normal density or skewed density, and z_p is median or extreme quantiles are investigated.

Standard normal kernel and uniform kernel are used for kernels K and W, and all smoothing parameters are chosen to minimize their asymptotic MISE, these involve calculation of $q_p'(x)$, $q_p''(x)$, $\chi(z)$ and $g(x)$. In method N0.7, the asymptotic MISE does not depend on parameter k, and in simulations different k has little effect on ISE, average of ISE is summarized according by three different k values $(k = 10, 20, 30)$ for $n = 100$ and $(k = 20, 30, 50)$ for $n = 500$ are used.

The regression quantiles are estimated in the whole interval and in a subinterval, which result in different smoothing parameters for quantiles smoothing. The first three models are designed to have $X \sim N(0, \sigma^2)$ irrespective of high density or low density, and comparing estimated regression quantiles in a small interval of x and in a bigger interval $[min(X), max(X)]$ of x's range, while the last model is partitioned into $x \in [1, 4]$ and $x \in [0, 5]$ because of $x \sim U[0, 5]$. The first two models have normal random errors while the last two have exponential random errors.

Therefore, we have the following three comparisions: an overall one, one (hopefully) without at boundary influence, and one for boundaries only.

The following quantities are involved in the calculation of smoothing parameters:

$\mu_2(K) = 1$, $R(K) = 0.2820946$, $\mu_2(W) = \alpha(W) = 1/3$, $R(W) = 0.5$, $R(K_s) = \sqrt{2}/16$ where K_s is the equivalent kernel of cubic spline (Silverman, 1984): $K_s(u) = \frac{sin(|u|/\sqrt{2}+\pi/4)}{2} e^{-|u|/\sqrt{2}}$.

With the above mentioned frame in mind four special cases of model (7.1) are considered and properties of $q_p(x)$ are investigated empirically.

7.2 Simulation 1

Random samples of size $n = 100, 500$ are simulated from the model

$$Y = sin(0.75X) + 1 + 0.3\epsilon. \tag{7.3}$$

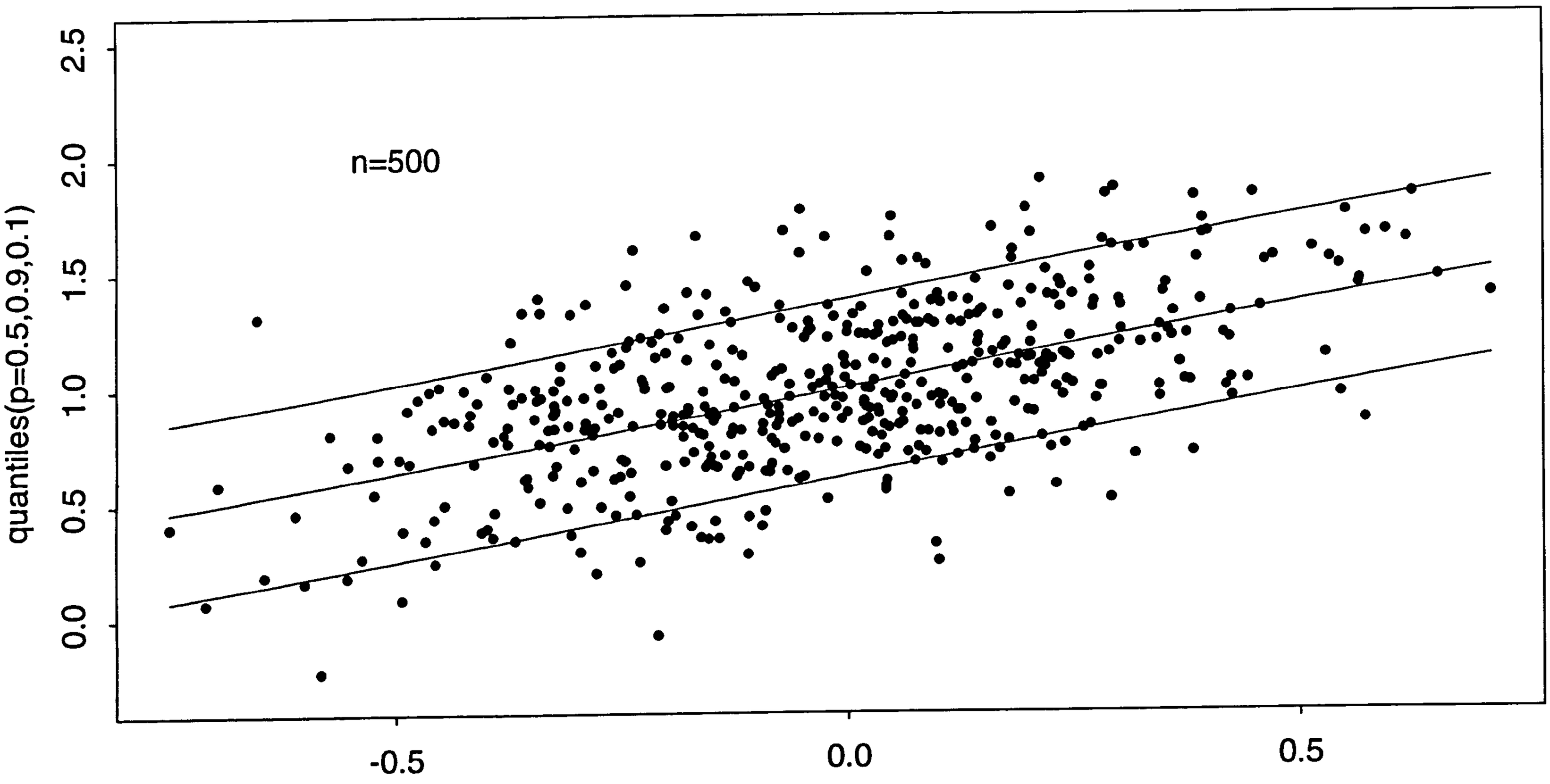
with $\epsilon \sim N(0,1)$ independent of $X \sim N(0, 0.25^2)$.

Then

$$\begin{aligned} q_p(x) &= sin(0.75x) + 1 + 0.3.\Phi^{-1}(p) \\ q_p'(x) &= 0.75cos(0.75x) \\ q_p''(x) &= -0.75^2 sin(0.75x) \end{aligned}$$

The curves of p-quantiles for $p = 0.1$, $p = 0.5$ and $p = 0.9$ are displayed in Figure 7.1(a).

The features of this model are almost linear quantile curves, normal density design and normal-type conditional distribution.



7.1: $Y=1+\sin(0.75X)+0.3e$, $X\sim N(0,0.0625)$, $e\sim N(0,1)$

i) ISE in Interior Points

The ISE of $\hat{q}_p(x)$ multiplied by 1000 for interior points are calculated for x in subinterval [-0.3, 0.3] of the whole interval [-0.58623294, 0.66856335] when n=100 and [-0.7087922, 0.7295326] when n=500 which are given in Tables 7.1 to 7.4.

It is seen that quantile estimators give better fitting in the subintervals than in whole interval calculated from $[minX, maxX]$ for either value of n: boundary effects are quite considerable.

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	1.45	2.15	2.23
N0.2	0.96	1.68	1.73
N0.3	0.92	1.34	1.34
N0.4	0.91	1.0	0.98
N0.5	1.0	1.3	1.48
N0.6	1.0	1.3	1.48
N0.7	1.0	0.75	0.96

Table 7.1: ISE based on Model 7.3 for n=100 in interval [-0.3,0.3]

For $n = 100$, all but N0.1 seem fairly comparable, although N0.2 & 3 and perhaps 5&6 seem to be worse than 4 & 7 at extremes. There are no great qualitative differences between interior only and overall. Also, for $n = 500$, all but No.1 seem fairly comparable in interior, but there remain differences between methods overall presumedly due to boundary effects (N0s 2, 5 &6 do "badly").

ii) AD in Boundary Points

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	10.8	21.65	23.7
N0.2	4.1	10.5	11.15
N0.3	3.34	6.2	6.1
N0.4	3.28	4.6	4.0
N0.5	4.1	9.3	9.88
N0.6	4.08	9.3	9.7
N0.7	3.97	6.2	6.0

Table 7.2: ISE based on Model 7.3 for n=100 in x's interval [-0.58623294, 0.66856335]

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	0.47	0.78	0.8
N0.2	0.198	0.35	0.29
N0.3	0.195	0.33	0.33
N0.4	0.194	0.22	0.22
N0.5	0.21	0.31	0.36
N0.6	0.21	0.30	0.31
N0.7	0.25	0.25	0.25

Table 7.3: ISE based on Model 7.3 for n=500 in interval [-0.3,0.3]

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	5.3	7.2	5.85
N0.2	1.56	13.9	4.24
N0.3	0.92	3.78	3.26
N0.4	0.89	1.93	1.49
N0.5	1.55	19.9	1.48
N0.6	1.56	20.5	5.7
N0.7	0.83	2.98	3.08

Table 7.4: ISE based on Model 7.3 for n=500 in x's interval [-0.7087922, 0.7295326]

The absolute deviations of the estimators $\hat{q}_p(x)$ when $g(x) = 4\phi(x/0.25)$ are calculated at left and right boundary points x_L and x_R respectively. For n=100, $x_L = -0.58623294$ and $x_R = 0.66856335$ while $x_L = -0.7087922$ and $x_R = 0.7295326$ for n=500. The values of AD, multiplied by 100, are given in Tables 7.5 & 7.6.

At the most extreme positions of X's interval, all but N0.1 seem fairly comparable, although N0.2, 3, 4 & 7 do better than N0.5 & 6. Also, for any method, the absolute deviations at left and right extreme points are not always consistently high or low.

Method		(L)			(R)	
p	0.5	0.9	0.1	0.5	0.9	0.1
N0.1	41.8	33.4	14.7	31	16	28.8
N0.2	7.1	9.6	19	7.7	2	1
N0.3	6	7.4	11	6.1	11.3	7.7
N0.4	5.8	5.9	7.1	5.8	7.6	7.15
N0.5	29	22.5	23.6	18.48	19	22.8
N0.6	29	19.8	23.5	18	19	22.5
N0.7	6.25	8.4	7.4	29.5	52.9	52.4

Table 7.5: AD in x's left ($L = -0.58623294$) and right ($R = 0.66856335$) boundary points based on Model 7.3 with n=100

Method		(L)			(R)	
p	0.5	0.9	0.1	0.5	0.9	0.1
N0.1	8.6	9	12.7	13.7	20	9.75
N0.2	5	9.4	14.7	4.7	33	5.6
N0.3	3.1	4	8	3.2	11.5	5.3
N0.4	3	3	5	2.8	4.7	4
N0.5	5.1	8.7	8.6	5.2	8	9.2
N0.6	5.1	9	9	5.1	6.8	8.5
N0.7	8	5.3	12.75	4.8	5.2	4.45

Table 7.6: AD in x's left ($L = -0.7087922$) and right ($R = 0.7295326$) boundary points based on Model 7.3 with n=500

7.3 Simulation 2

Instead of the linear-type quantile curves, now we hope to fit quantile curves with a combination of quadratic-arc and jump-type variations. Data are generated from the model (for n=100 and 500)

$$Y = 2.5 + sin(2X) + 2exp(-16X^2) + 0.5\epsilon \tag{7.4}$$

where $X \sim N(0,1)$ is higher-density design in central region than that of model 1 and $\epsilon \sim N(0,1)$ too. Then

$$q_p(x) = 2.5 + sin(2x) + 2e^{-16x^2} + 0.5\Phi^{-1}(p) \tag{7.5}$$

$$q_p'(x) = 2cos(2x) - 2^6 xe^{-16x^2} \tag{7.6}$$

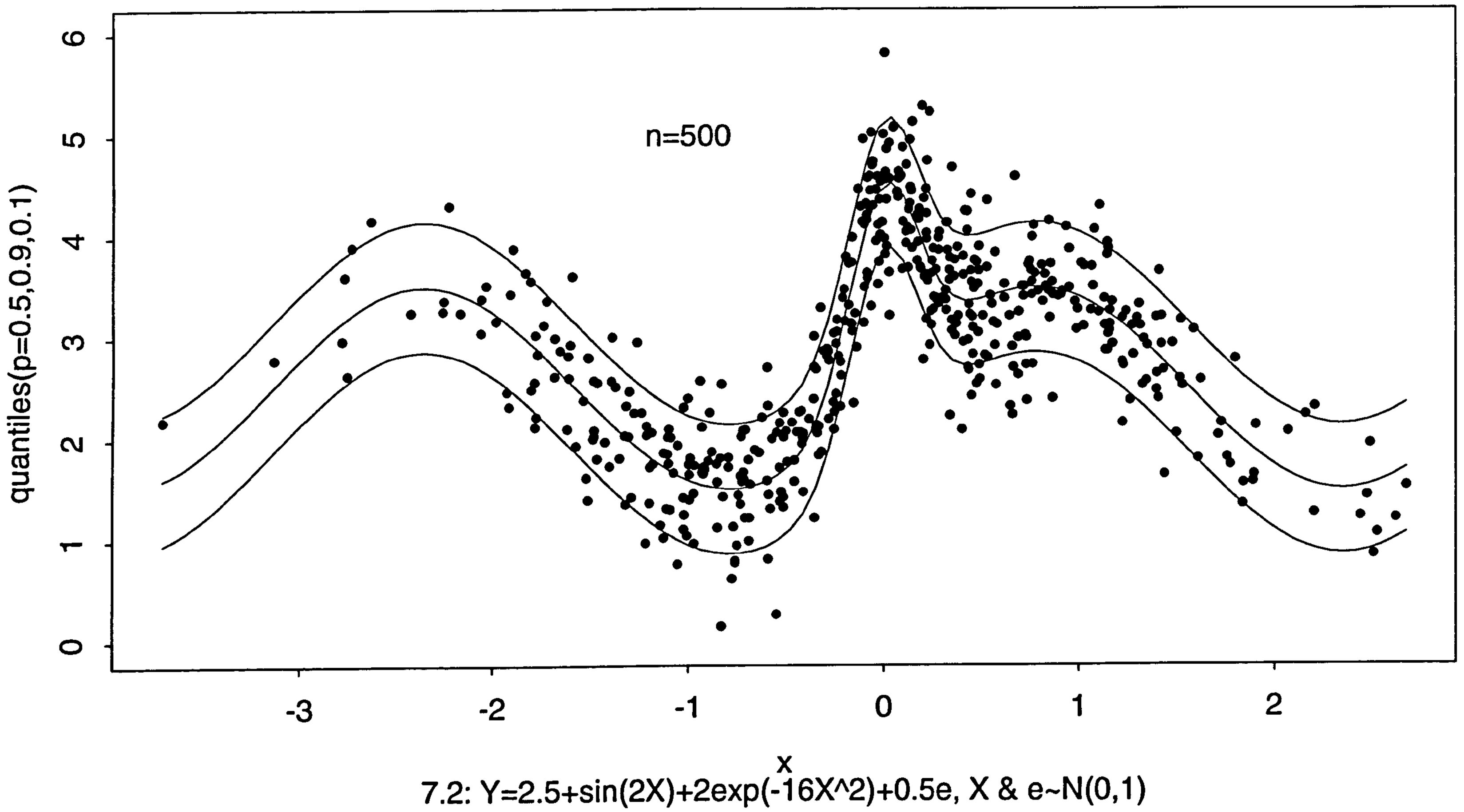
$$q_p''(x) = -4sin(2x) + 2^{11}x^2e^{-16x^2} - 2^6e^{-16x^2} \tag{7.7}$$

The quantile curves $q_p(x)$ are shown in Figure 7.1(b) for $p = 0.1, 0.5$ and 0.9.

The ISE multiplied by 1000 for the estimators $\hat{q}_p(x)$ and for interior point x in the subinterval [-1,1] and the whole interval are given in Table 7.7 and 7.8 when n=100 and Table 7.9 and 7.10 when n=500.

Similarly the AD multiplied by 100 are calculated at $X_L = -2.473947$ and $x_R = 2.83873$ for n=100, while $x_L = -3.368479$ and $x_R = 2.930744$ for n=500 respectively, and these are displayed in Table 7.11 and 7.12.

At interior points with overall measurement, two local constant kernel fittings N0.1 & 3 seem do better than others for $n = 100$, while all seven methods are comparable for $n = 500$. However, once increasing the range of X, local linear fitting (N0.2, 4 & 7) are superior. Also, semi-parametric methods (N0.5 &6) seem to worse in the bigger intervals in terms of overall measurement.



7.2: Y=2.5+sin(2X)+2exp(-16X^2)+0.5e, X & e~N(0,1)

At boundary points, for $n = 100$, N0.7 does best while N0.1 & 2 do worse for $p = 0.5$, and N0.3 & 5 have smaller AD than others for $p = 0.9\&0.1$. However, for $n = 500$, all but N0.1 are comparable for $p = 0.5$, while N0.3 & 4 do best for $p = 0.9\&0.1$.

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	58.8	84.1	85.3
N0.2	63.7	86.8	86.9
N0.3	58.7	80.8	81
N0.4	63.7	86.6	86.9
N0.5	65.8	93.2	91.3
N0.6	64.3	93.5	90.3
N0.7	64	89.4	88.4

Table 7.7: ISE based on Model 7.4 for n=100 in x's interval [-1, 1]

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	392.77	693.84	471.4
N0.2	367.98	590.5	444
N0.3	420.7	343	388.5
N0.4	395.2	246	286.2
N0.5	450.3	478.1	443.2
N0.6	443.8	469.1	471
N0.7	357.1	528.4	444.6

Table 7.8: ISE based on Model 7.4 for n=100 in x's interval [-2.473947, 2.83873]

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	18.66	22.36	21.47
N0.2	18.27	21.43	20.67
N0.3	18.65	21.37	20.44
N0.4	18.24	21.32	20.35
N0.5	18.52	21.98	21.4
N0.6	18.5	21.68	21.04
N0.7	19.37	19.64	20.76

Table 7.9: ISE based on Model 7.4 for n=500 in x's interval [-1, 1]

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	122.1	573.1	827.2
N0.2	49.4	661.7	615.7
N0.3	49.4	184.7	184.7
N0.4	47.6	121.7	164.5
N0.5	70.5	203.2	203.4
N0.6	70.4	201.1	195.9
N0.7	48.6	189	173.2

Table 7.10: ISE based on Model 7.4 for n=500 in x's interval [-3.368479, 2.930744]

Method		(L)			(R)	
p	0.5	0.9	0.1	0.5	0.9	0.1
N0.1	35.47	23.54	39.1	40	38.3	61.4
N0.2	17.56	37.2	68.5	20.4	15.6	10.4
N0.3	20.1	15.4	15.2	19.8	20	18.1
N0.4	20.9	23.2	28.6	19.03	29	19.3
N0.5	21.3	21	19.7	19.9	20	18.7
N0.6	20.3	22.3	20.1	20.1	21.9	21
N0.7	16.4	34.2	26.5	17.6	20.2	21.1

Table 7.11: AD in boundary points based on Model 7.4 for n=100

Method		(L)			(R)	
p	0.5	0.9	0.1	0.5	0.9	0.1
N0.1	15.6	21.7	72	26.7	22.2	21
N0.2	8.3	69	66	14.2	36.4	54.5
N0.3	18.3	19.4	19.4	14.2	20.7	20.7
N0.4	7.88	19.2	11.9	14.2	19.9	20.8
N0.5	15.9	46.7	58	15.5	39.9	49.8
N0.6	16.9	50.3	53.2	16.5	51.2	54.2
N0.7	13.4	30.8	26.8	15	35.1	25.6

Table 7.12: MSE in boundary points based on Model 7.4 for n=500

7.4 Simulation 3

A skew distribution model of the form (7.8) is fitted here.

$$Y = 2 + 2cos(X) + exp(-4X^2) + \epsilon \tag{7.8}$$

where $X \sim N(0,1)$ and $\epsilon \sim E(1)$. So

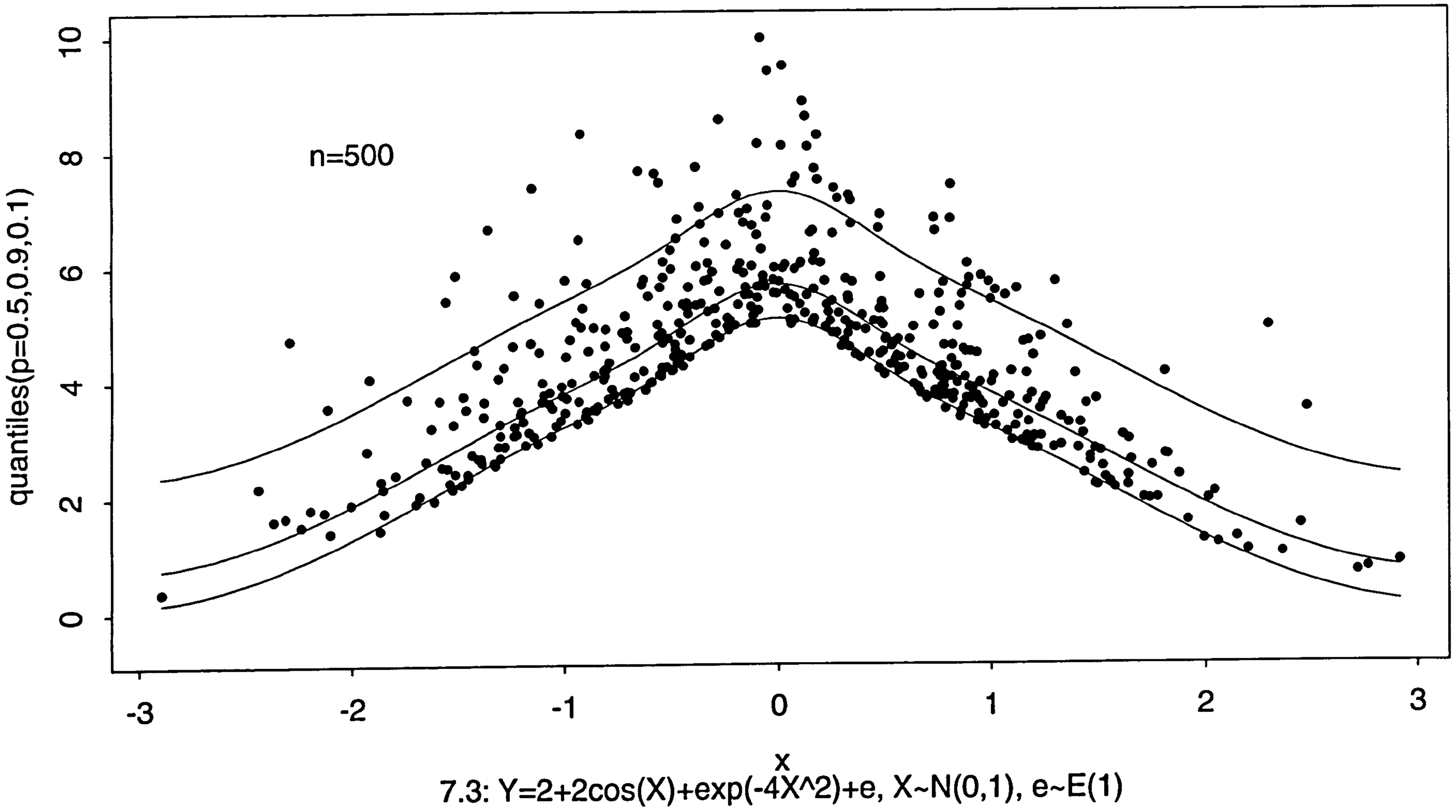
$$\begin{aligned} m(x) &= 2 + 2cos(x) + e^{-4x^2} + 1 \\ q_p(x) &= 2 + 2cos(x) + e^{-4x^2} - log(1-p) \\ q_p'(x) &= -2sin(x) - 8xe^{-4x^2} \\ q_p''(x) &= -2cos(x) - 8e^{-4x^2} + 64x^2e^{-4x^2} \end{aligned}$$

The curves of $q_p(x)$ for $p = 0.1, 0.5$ and 0.9 are shown in Figure 1(c), estimators $\hat{q}_p(x)$ are fitted and their ISE calculated for n=100 and 500 at interior points x in the subinterval [-1, 1] and the whole interval [-2.3, 2.1]. The results multiplied by 1000 are displayed in Tables 7.13 to 7.16. The reason of taking [-2.3, 2.1] as whole interval for both n=100 and 500 is that there are almost no data points out of this range.

As for comparison at boundary points, AD are calculated when $g(x) = \phi(x)$ and for $x_L = -2.3$ and $x_R = 2.1$. The results are displayed in Tables 7.17 and 7.18.

At interior intervals with n=100 & 500, all are comparable for median, and N0.3 & 4 do best for $p = 0.9$ wile N0. 1 & 2 do best for $p = 0.1$. However, in the bigger interval [-2.3, 2.1], all but N0.1, 3 & 7 are comparable for median, while N0.7 does best for extreme quantiles.

For $n = 100$, all methods do better at left boundary point $x = -2.3$ than at right boundary point $x = 2.1$. Contrary to this, all methods do better at $x = 2.1$ than at $x = -2.3$, where n=500.



7.3: Y=2+2cos(X)+exp(-4X^2)+e, X~N(0,1), e~E(1)

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	151.7	208.3	150.7
N0.2	146.7	205	153
N0.3	147.9	168	142.8
N0.4	140.5	162.4	176.8
N0.5	146.8	196.7	193.8
N0.6	146.8	192.3	174.7
N0.7	152	172.1	183.1

Table 7.13: ISE based on Model 7.8 for n=100 in [-1, 1]

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	36.5	57.9	53.1
N0.2	33.7	56.6	60.3
N0.3	33.6	39.7	100
N0.4	33	36.5	140
N0.5	34.7	52.5	89.5
N0.6	34	51.5	96.5
N0.7	35.3	53.2	56.5

Table 7.14: ISE based on Model 7.8 for n=500 in [-1,1]

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	566.2	687.2	693
N0.2	399.9	547.1	623.1
N0.3	457.7	634.1	654.4
N0.4	346.5	538.9	675.1
N0.5	392.1	593.1	693.1
N0.6	393	555.5	562.2
N0.7	504.8	510	510

Table 7.15: ISE based on Model 7.8 for n=100 in [-2.3,2.1]

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	149.2	244.9	276.8
N0.2	103.5	221.8	268.9
N0.3	148.9	230	270.5
N0.4	112	210.8	227.7
N0.5	132.1	240.1	257.7
N0.6	125.1	225.7	234.1
N0.7	140.1	193	193.5

Table 7.16: ISE based on Model 7.8 for n=500 in [-2.3,2.1]

Method		(L)			(R)	
p	0.5	0.9	0.1	0.5	0.9	0.1
N0.1	95.19	156.38	167.7	43.51	48.3	50.2
N0.2	165	134.2	133.3	40.3	46.03	34.2
N0.3	145	135.2	143.3	33.6	44.3	46.2
N0.4	169	98.6	108.9	36	39.5	57.9
N0.5	98.5	162.1	156.4	43.2	46.1	98.7
N0.6	104.6	133.5	45.7	23.1	49.9	32.1
N0.7	27.5	69.7	33.4	24.3	67.1	53.1

Table 7.17: AD in boundary based on Model 7.8 with n=100

Method		(L)			(R)	
p	0.5	0.9	0.1	0.5	0.9	0.1
N0.1	18.3	46.7	51.3	148.5	132.4	132.8
N0.2	28.7	20.2	25	173.4	0.127.4	161
N0.3	41.3	23.2	42.1	188.7	156.4	179.6
N0.4	44.4	55.5	47.6	194.1	145.1	167
N0.5	34.6	16.9	23.1	106.3	176.5	134.2
N0.6	22.6	24.1	34.6	144.3	143.2	97.6
N0.7	20.5	27.1	19.1	98.5	198.5	196.5

Table 7.18: MSE in boundary based on Model 7.8 with n=500

7.5 Simulation 4

Consider a heteroscedastic model:

$$Y = 2 + X + exp(-X)\epsilon \tag{7.9}$$

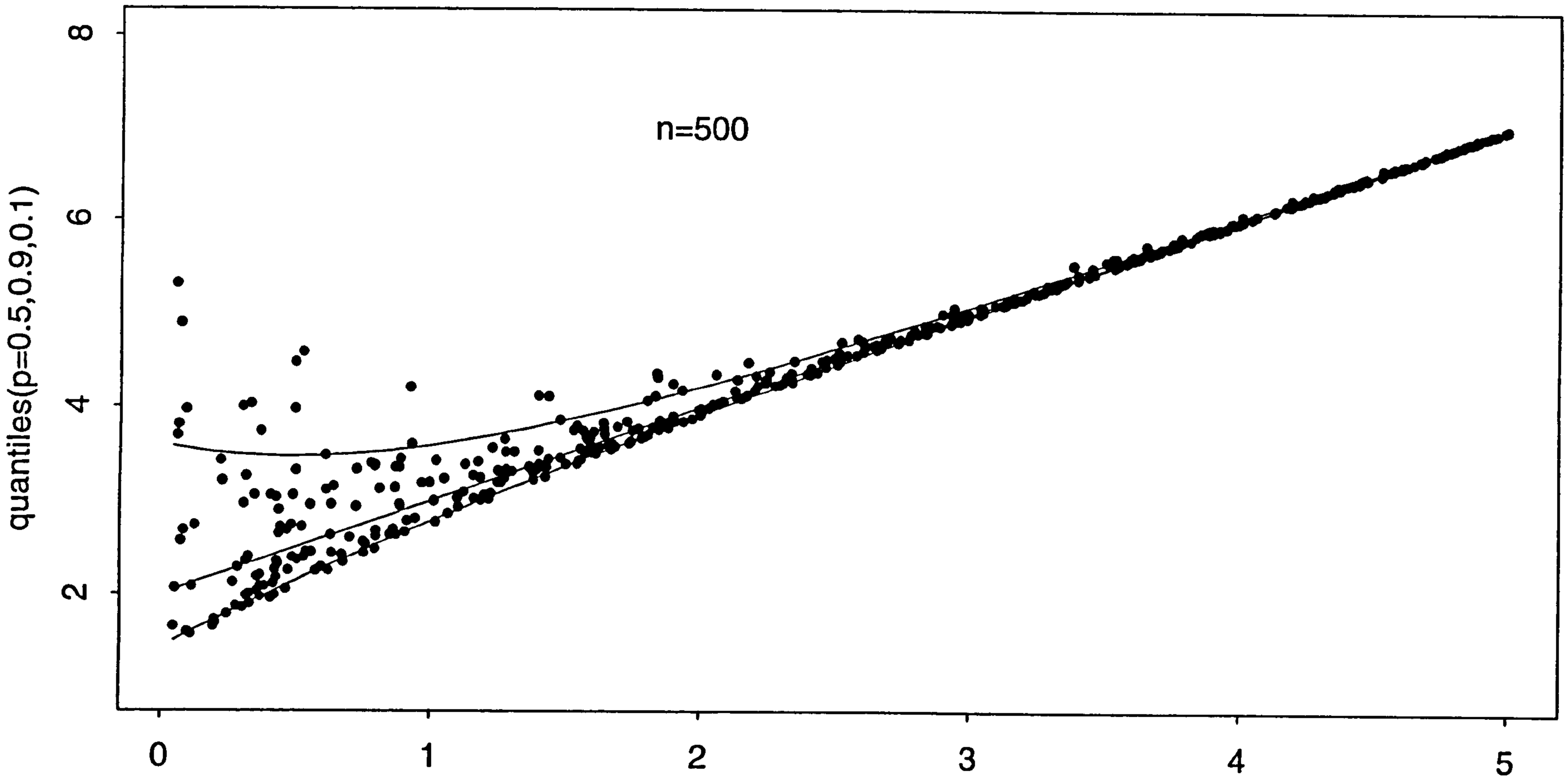
with $\epsilon \sim E(1) - log(2)$, and $X \sim U(0,5)$, then

$$\begin{aligned} m(x) &= 2 + x + e^{-x}(1 - log(2)) \\ q_p(x) &= 2 + x - e^{-x}log(2(1-p)) \\ q_p'(x) &= 1 + e^{-x}log(2(1-p)) \\ q_p''(x) &= -e^{-x}log(2(1-p)) \end{aligned}$$

Figure 1 (d) shows the graphs of 10th, 50th and 90th quantiles for this model.

The ISE, multiplied by 1000, for n=100 and 500 at the interior points are given in Table 7.19 to 7.22 for $x \in [1,4]$ and $x \in (0,5)$ respectively. At the boundary points assume that $X \sim U(0,5)$. As seen from Figure 1(d) there is a possibility of large bias at the left boundary, thus taking $x_L = 0.05$ as $g(x)$'s left boundary point and $x_R = 4.95$ as the boundary point then calculate the AD's. The results multiplied by 100 are displayed in Table 7.23 and 7.24 respectively.

At interval [1, 4], N0.2, 3 & 4 do best for $n = 100\&500$, while all but N0.1 are comparable in interval (0, 5). Similarly, at boundary points, all but N0.1 seem comparable.



7.4: Y=2+X+exp(-X)e, X~U(0,5), E~E(1)-log(2)

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	8.45	56.78	6.84
N0.2	3.36	32.37	3.15
N0.3	1.01	8.58	7.16
N0.4	4.15	4.62	1.1
N0.5	8.9	59.9	79.1
N0.6	8.8	43.3	72.1
N0.7	8	18.9	21.1

Table 7.19: ISE based on Model 7.9 in [1, 4] with n=100

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	1.62	15.37	2.1
N0.2	0.69	7.72	0.54
N0.3	0.1	2.2	2.19
N0.4	0.88	1.7	0.65
N0.5	1.4	4.5	6.98
N0.6	1.53	3.4	7
N0.7	5.7	4.8	7.3

Table 7.20: ISE based on Model 7.9 in [1, 4] with n=500

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	60.89	106.6	24.4
N0.2	32	17.9	21.5
N0.3	53.2	65.3	32.15
N0.4	30.65	13.22	21.68
N0.5	39.58	32.19	32.1
N0.6	36.1	31.2	32.1
N0.7	31.1	67.4	43.2

Table 7.21: ISE based on Model 7.9 in whole interval (0, 5) with n=100

Method	$p = 0.5$	p=0.9	p=0.1
N0.1	34.45	28.86	16.42
N0.2	5.39	17.3	7.1
N0.3	19.1	16.7	20.5
N0.4	8.17	6.78	5.46
N0.5	10.3	10.3	10
N0.6	19.68	10.49	13.2
N0.7	4.88	48.48	3.1

Table 7.22: ISE based on Model 7.9 in whole interval (0, 5) with n=500

Method		(L)			(R)	
p	0.5	0.9	0.1	0.5	0.9	0.1
N0.1	21.55	18.51	9.67	35.47	38.5	19.3
N0.2	20.9	4.7	12.75	0.32	13.2	0.68
N0.3	11.2	13.57	13.9	26.98	27.06	30.1
N0.4	5	11.78	11.43	3.77	0.61	3.68
N0.5	3.21	12.9	13.2	11.2	9.85	12.69
N0.6	12.9	9.4	10.56	5.43	21.48	9.85
N0.7	16.47	56.3	13	2.2	3	10.95

Table 7.23: AD in boundary 0.05 and 4.95 based on Model 7.9 with n=100

Method		(L)			(R)	
p	0.5	0.9	0.1	0.5	0.9	0.1
N0.1	24.9	13.15	17.65	28.28	15.15	17.15
N0.2	9.54	13.87	18.25	0.13	0.96	0.71
N0.3	20.18	18.52	9.7	19.83	18.4	0.15
N0.4	14.68	9.7	9.7	4	0.15	0.15
N0.5	8.6	9	12.7	13.7	19.9	9.76
N0.6	5	9.48	14.68	4.7	32.94	5.6
N0.7	7.8	2.7	5	1.68	7.34	10

Table 7.24: AD in boundary points 0.05 and 4.95 based on Model 7.9 with n=500

7.6 Discussion of the Results

The numerical results of the previous section as presented draw the following remarks.

1) The boundary area keeps large ISE as compared to ISE in the subinterval, as ISE of the whole interval is at least twice as that of its subinterval. The local linear methods do not make significant decrease as local constant methods while shifting interval from big to small one.

2) As sample size n increases, all ISE significantly decrease irrespective of the interval width, and this is particularly significant for normal error model (No.1 and No.2).

3) In terms of ISE, double-kernel methods have overall best perform and the local constant double-kernel (No.3) often perform better than local linear single-kernel method (No.2) under normal error model (No.1 and No.2), but not for exponential error model (No.3 and No.4).

4) For all methods, ISE's of the extreme quantiles are usually bigger than that of median point, these differ from each other, no pattern is noticeable when comparing 10th and 90th quantiles.

5) Absolute deviation (AD) at the boundary varies irrespective of the value of n, that is, for fixed points near the boundary, no one method is always preferable in terms of AD. The better methods are N0.4, 5 & 7.

6) Theoretically, when low design density is specified, local linear fitting is expected to have significant good performance, but simulation of model 7.3 and

7.4 show that there is not too big difference with high design density model. On the other hand, in the subinterval [-1, 1] of model 7.4, local constant fitting is a little superior to local linear fitting. Also, theoretical investigation supports that the local constant fitting and local linear fitting have almost same performance when design density is uniform, which is not supported totally by the empirical simulation model 7.4.

In all, taking account of all factors such as optimality in terms of ISE, boundary performance and computing time save, N0.4, 5 & 7 should be recommended to use in practice.

Chapter 8

Local Polynomial Fit with kth Derivative Penalized Least-Squares Smoothing

8.1 Introduction

Nonparametric regression approach for estimating a curve f given observations $Y_i = f(t_i) + \varepsilon_i$ provides a useful diagnostic tool for data analysis. It is proposed to investigate the estimation of the regression function and its derivatives at a point by "local" fitting of an mth degree polynomial to the data via kth derivative penalized least-squares smoothing. The structure of the derivative penalized least-squares smoothing is discussed first. Section 8.3 carries out the calculation for local linear fit with 2nd derivative penalized, while Section 8.4 illustrates the approximate equivalent weightings. Section 8.5 is concerned with the equivalent kernel and the boundary properties of this method. Finally, a section considers

the integrated mean squared error and provides an outlook to future work.

8.2 The Structure of Local Polynomial Fit with Derivative Penalized

Consider the problem of estimating the function $f(t)$ from noisy observation

$$Y_i = f(t_i) + \varepsilon_i \tag{8.1}$$

where the errors ε_i are iid random variables with finite variance σ^2 and the design points t_i are equispaced in [0, 1]. Observations (t_i, y_i) $i = 0, 1, ..., n$ are given and assume that

$$t_0 < t_1 < ... < t_n$$

It is known that a spline estimate $g(t)$ of $f(t)$ is usually given as the minmizer over function g of

$$\frac{1}{n}\sum_{0}^{n}(y_i - g(t_i))^2 + \lambda \int_0^1 g^{(k)}(t)^2 dt$$

this is here called local constant fit with kth derivative least-squares smoothing and the cubic spline with $k = 2$ is particularly popular.

Now, the idea of the local polynomial fit is that the estimator $g(\cdot)$ of $f(\cdot)$ is obtained by fitting polynomial function

$$\begin{aligned} g(z) &= g(t) + g^{(1)}(t)(z-t) + \frac{1}{2}(z-t)^2 g^{(2)}(t) + \cdots + \frac{1}{m!}(z-t)^m g^{(m)}(t) \\ &\equiv g(t) + g_1(t)(z-t) + \frac{1}{2}(z-t)^2 g_2(t) + \cdots + \frac{1}{m!}(z-t)^m g_m(t) \end{aligned} \tag{8.2}$$

to the (t_j, y_j) $(j = 0, 1, ..., n)$ for z in a neighborhood of t. Here we fit the local polynomial by using derivative penalized least-squares smoothing, where m is the

degree of fitted polynomial, and when $m = 0$, it just corresponds to local constant fitting, the usual spline smoothing.

Locally, estimating $f(t)$ is still equivalent to obtaining $g(t)$. At the same time, estimating $f^{(1)}(t)$ is equivalent to obtaining $g_1(t)$, $f^{(m)}(t)$ to $g_m(t)$. This motivates us to define an estimator by setting $\hat{f}(t) = g(t)$, where $g(t), g_1(t), g_2(t), \cdots, g_m(t)$ minimize

$$\begin{aligned} & S \quad (g, g_1, ..., g_m) = \\ & \textstyle\sum_0^n \ (y_j - g(t_j) - (t - t_j)g_1(t_j) - \frac{1}{2}(t - t_j)^2 g_2(t_j) - \cdots - \frac{1}{m!}(t - t_j)^m g(t_j))^2 \\ & + \quad R(g, g_1, \cdots, g_m) \end{aligned} \tag{8.3}$$

where $R(g, g_1, \cdots, g_m)$ is the functional which quantifies the roughness of $g, g_1, \cdots, g_m$. A natural and convenient choice of $R(g, g_1, \cdots, g_m)$ is

$$R(g, g_1, \cdots, g_m) = \lambda \int_0^1 g^{(k)}(t)^2 dt + \lambda_1 \int_0^1 g_1^{(k)}(t)^2 dt + \cdots + \lambda_m \int_0^1 g_m^{(k)}(t)^2 dt \tag{8.4}$$

where smoothing parameters λ, λ_1, ..., λ_m are non-negative.

Alternatively, one may specify the jth derivative of $g(t)$ as g_j $(j = 1, ..., m)$ of $R(g, g_1, \cdots, g_m)$. For example, consider local linear fit with 2nd derivative penalized (penalized on g) least-squares smoothing: rewrite $S(g, g_1)$ as $S(g)$:

$$S(g) = \sum_0^n (y_j - g(t_j) - (t - t_j)g^{(1)}(t_j))^2 + \beta \int_0^1 g^{(2)}(t)^2 dt + \beta_1 \int_0^1 g^{(3)}(t)^2 dt \tag{8.5}$$

g is the minimizer of $S(g)$, so we must have $S(g + \alpha h) \geq S(g)$ for any $h \in C^3$ and $\alpha \in R$.

$$S(g + \alpha h) - S(g) =$$

$$\sum_0^n (y_j - g(t_j) - (t - t_j)g^{(1)}(t_j) - \alpha h(t_j) - \alpha(t - t_j)h^{(1)}(t_j))^2 - \sum_0^n (y_j - g(t_j) - (t - t_j)g^{(1)}(t_j))^2$$

$$+\beta\left[\int_0^1 (g^{(2)}(t) + \alpha h^{(2)}(t))^2 dt - \int_0^1 g^{(2)}(t)^2\right] + \beta_1\left[\int_0^1 (g^{(3)}(t) + \alpha h^{(3)}(t))^2 dt - \int_0^1 g^{(3)}(t)^2\right] \geq 0$$

Expanding the expression above and since this holds for all $\alpha \in R$, we must have the coefficient of α zero for any $h \in C^3$. This means

$$\sum_j (h(t_j) + (t - t_j)h^{(1)}(t_j))(y_j - g(t_j) - (t - t_j)g^{(1)}(t_j))]$$
$$= \beta \int_0^1 h^{(2)}(t)g^{(2)}(t)dt + \beta_1 \int_0^1 h^{(3)}(t)g^{(3)}(t)dt \tag{8.6}$$

Taking j fixed and $h(t)$ to be,

$$h(t) = \begin{cases} (t - t_j)^4(t - t_{j+1})^4 & \text{if } t_j \le t \le t_{j+1} \\ 0 & \text{otherwise} \end{cases} \tag{8.7}$$

Clearly, from $h(t_j) = 0$

$$\beta \int_0^1 h^{(2)}(t)g^{(2)}(t)dt + \beta_1 \int_0^1 h^{(3)}(t)g^{(3)}(t)dt = 0 \tag{8.8}$$

Integrating by parts three times, it follows that

$$\beta g^{(4)}(t) - \beta_1 g^{(6)}(t) = 0 \tag{8.9}$$

for $\forall t \in [0, 1]$ except possibly t_j.

Solving the above equation,

$$g(t) \equiv c_1 \left((\frac{\beta}{\beta_1})^2 \right) e^{\pm\sqrt{\frac{\beta_1}{\beta}}t} + \frac{c_2}{6}t^3 + \frac{c_3}{2}t^2 + c_4 t + c_5 \tag{8.10}$$

on each $[t_j, t_{j+1}]$, where constants c_i $(i = 1, 2, ..., 5)$ are determined by $g^i(t_j)$ $(i = 1, 2, ..., 5)$.

This shows that g is locally the combination of an exponential function and a cubic function on each $[t_i, t_{i+1}]$. Corresponds to local constant fitting with 1st and 2nd derivative penalties on g, this approach may be better than simply a cubic spline when approximating g. Generally, under the consideration of $g^{(j)}(t)$ as $g_j(t)$ in $R(g, g_1, ..., g_m)$, g is the combination of a piecewise exponential function

and a piecewise polynomial of order k. (Note that (8.5) is a double smooth-parameter technique of spline version, and the conclusion of (8.10) is still true if keeping local constant fitting but with penalty term (8.5), and this seems to suggest that local polynomial fitting may not make much difference, but depends on the number of smoothing parameters). This would be an interesting result and we would like to leave it to be discussed in future somewhere and here regard only the roughness penalties as a sum of integrals of several derivatives of independent functions as equation (8.4). However, it will estimate nothing if using $R(g) = \lambda \int_0^1 g^{(k)}(t)^2 dt$ instead of $R(g, g_1, \cdots, g_m)$ which only penalizes regression function itself (Section 8.4). On the other hand, the estimator $g(t)$ defined from residual sum of squares always is related to $g_j(t)$ $(j = 1, \cdots, m)$ and $\lambda, \lambda_1, \cdots, \lambda_m$ which are called smoothing parameters and represent the rate of exchange between residual error and local variation, so only when all g_j are penalized the best compromise for $g(t)$ between smoothness and goodness-of-fit is attained.

8.3 The Algorithm for Local Linear Fit with 2nd Derivative Penalty

Like the local constant fit with 2nd derivative penalty, the local linear fit with 2nd penalty is still the most representative among the local polynomial fit with kth derivative penalized least-squares smoothing.

Now, the problem is of finding $g(t)$ and $g_1(t)$ for given $t \in$ [a, b] which minimizes

$$S(g, g_1) = \sum_{i=1}^{n} (y_i - g(t_i) - (t - t_i) g_1(t_i))^2 + \lambda \int_a^b g^{(2)}(t)^2 dt + \lambda_1 \int_a^b g_1^{(2)}(t)^2 dt \quad (8.11)$$

Theorem 8.3.1: Suppose $t_1, t_2, \cdots, t_n$ are real numbers on some interval [a, b]

satisfying $a < t_1 < t_2 < \cdots < t_n < b$. Then both $g(t)$ and $g_1(t)$ that minimize $S(g, g_1)$ are cubic splines and the points t_i are knots.

This conclusion is very natural one, as the solution to the interpolation problem is the function minimizing

$$\int_0^1 (g^{(2)}(t))^2 dt$$

subject to $g(t_i) = y_i$ $(i = 0, 1, \cdots, n)$.

By using the Reinsch method (1967), through some tedious work, the optimal functions $g(t)$ and $g_1(t)$ are determined from the corresponding Euler-Lagrange equations:

$$g^{(4)}(t) = 0 \text{ and } g_1^{(4)}(t) = 0,\ t_i < t < t_{i+1},\ i = 1, \cdots, n-1, \tag{8.12}$$

Given $t \in [a, b]$,

$$\begin{aligned} \lambda(g^{(3)}(t_i)_- - g^{(3)}(t_i)_+) &= g(t_i) + (t - t_i)g_1(t_i) - y_i \\ \lambda_1(g_1^{(3)}(t_i)_- - g_1^{(3)}(t_{i_+}) &= (t - t_i)(g(t_i) + (t - t_i)g_1(t_i) - y_i) \end{aligned} \tag{8.13}$$

This shows that there are possible jumps of the 3rd derivative of g and g_1 at $t = t_i$, and the jump of g_1 is $\frac{\lambda}{\lambda_1}(t - t_i)$ times that of g and their first two derivatives are continuous. This completes the proof of the theorem.

For sake of a uniform notation, impose *natural boundary conditions* on g and g_1, that is, their second and third derivatives are zero at a and b, thus, g and g_1 can be regarded as *natural cubic splines* (NCS) and specify $g(t)$ and $g_1(t)$ as

$$\begin{aligned} g(t) &= d_i(t - t_i)^3 + c_i(t - t_i)^2 + b_i(t - t_i) + a_i \text{ for } t_i < t < t_{i+1} \\ g_1(t) &= \overline{d_i}(t - t_i)^3 + \overline{c_i}(t - t_i)^2 + \overline{b_i}(t - t_i) + \overline{a_i} \text{ for } t_i < t < t_{i+1} \end{aligned} \tag{8.14}$$

Given constants $a_i, b_i, c_i, d_i, \overline{a_i}, \overline{b_i}, \overline{c_i}, \overline{d_i}$, $i = 0, \cdots, n$; define $t_0 = a$ and $t_{n+1} = b$. The NCS implies that $d_0 = \overline{d_0} = c_0 = \overline{c_0} = d_n = \overline{d_n} = c_n = \overline{c_n} = 0$, so that g

and g_1 are linear on the two extreme intervals $[a, t_1]$ and $[t_n, b]$. Further, a short manipulation as that of Reinsch (1967) yields

$$Q^T \mathbf{g} = R\mathbf{r} \tag{8.15}$$

and

$$Q^T \mathbf{g_1} = R\mathbf{r_1} \tag{8.16}$$

where vectors $\mathbf{g}$, $\mathbf{g_1}$, $\mathbf{r}$ and $\mathbf{r_1}$ are $(g_1, \cdots, g_n)$, $(g_{11}, \cdots, g_{1n})$, $(r_2, \cdots, r_{n-1})$ and $(r_{12}, \cdots, r_{1n-1})$, respectively and $g_i = g(t_i), g_{1i} = g_1(t_i)$, $r_i = g^{(2)}(t_i)$, and $r_{1i} = g_1^{(2)}(t_i)$. The matrices Q and R are band matrices with bandwidth 3, and R is strictly diagonal dominant (strictly positive-definite) and tridiagonal matrix.

Let $h_i = t_{i+1} - t_i$ for $i = 1, \cdots, n-1$, then Q is the $n \times (n-2)$ matrix with entries q_{ij}, for $i = 1, \cdots, n$ and $j = 2, ..., n-1$,

$$q_{j-1,j} = h_{j-1}^{-1},\ q_{jj} = -h_{j-1}^{-1} - h_j^{-1},\ \text{ and }\ q_{j+1,j} = h_j^{-1} \tag{8.17}$$

for $j = 2, \cdots, n-1$, and $q_{ij} = 0$ for $\mid i - j \mid \geq 2$. The symmetric matrix R is $(n-2) \times (n-2)$ with elements r_{ij} for i and $j = 2, ..., (n-1)$, given by

$$\begin{aligned} r_{ii} &= \frac{1}{3}(h_{i-1} + h_i) \text{ for } i = 2, \cdots, n-1, \\ r_{i,i+1} &= r_{i+1,i} = \frac{1}{6}h_i \text{ for } i = 2, \cdots, n-2, \end{aligned} \tag{8.18}$$

and $r_{ij} = 0$ for $\mid i - j \mid \geq 2$. and

This result for g (g_1) is similar to that of Green and Silverman (1994) and it is stated as a theorem.

Theorem 8.3.2: The vectors $\mathbf{g}$ and $\mathbf{r}$ specify a NCS g if and only if the condition

$$Q^T \mathbf{g} = R\mathbf{r} \tag{8.19}$$

If the condition above is satisfied then the roughness penalty will satisfy

$$\int_a^b g^{(2)}(t)^2 dt = \mathbf{r}^T R\mathbf{r} = \mathbf{g}^T K\mathbf{g} \tag{8.20}$$

where

$$K = QR^{-1}Q^T \tag{8.21}$$

Now, to express $S(g, g_1)$ in terms of these vectors and matrices, let $\mathbf{Y}$ be the vector $(Y_1, \cdots, Y_n)^T$, and W=$diag(t - t_1, \cdots, t - t_n)$ for given $t \in [a, b]$, then rewriting $S(g, g_1)$ as

$$\begin{aligned} S(g, g_1) &= \mathbf{g}^T(I + \lambda K)\mathbf{g} + \mathbf{g_1}(W^2 + \lambda_1 K)\mathbf{g_1} \\ &- 2Y^T\mathbf{g} - 2Y^T W\mathbf{g_1} + 2\mathbf{g}^{\mathbf{T}} W\mathbf{g_1} + Y^T Y \end{aligned} \tag{8.22}$$

Since K is non-negative definite, then both matrices $I + \lambda K$ and $W^2 + \lambda_1 K$ are strictly positive-definite. It follows that the above equation has a unique minimum obtained by solving simultanously the following two linear equations for g and g_1:

$$\begin{aligned} (I + \lambda K)\mathbf{g} + W\mathbf{g_1} &= Y \\ (W^2 + \lambda_1 K)\mathbf{g_1} + W\mathbf{g} &= WY \end{aligned} \tag{8.23}$$

Clearly, if either $\mathbf{g_1}$ or $\mathbf{g}$ is known, then the solution to $\mathbf{g}$ and $\mathbf{g_1}$ is obtained using local constant fit cubic smoothing spline program (Green and Silverman, 1994). Rewriting above equations as

$$(I + \lambda K)\mathbf{g} = Y - W\mathbf{g_1} \tag{8.24}$$

$$(W^2 + \lambda_1 K)\mathbf{g_1} = W(Y - \mathbf{g}) \tag{8.25}$$

substituting for $\mathbf{g_1}$ or for $\mathbf{g}$ in either of the equations above it follows that

$$[(I + \lambda K) - W(W^2 + \lambda_1 K)^{-1}W]\mathbf{g} = [I - W(W^2 + \lambda_1 K)^{-1}W]Y \tag{8.26}$$

or

$$[W^2 + \lambda_1 K - W(I + \lambda K)^{-1} W]\mathbf{g_1} = W[I - (I + \lambda K)^{-1}]Y \tag{8.27}$$

which can be solved by Cholesky decomposition of the coefficient matrices as in Green and Silverman (1994).

8.4 Asymptotic Theory: The Approximate Equivalent Weightings

It is well known (cf. Wahba, 1975) that the spline smoother of local constant fit with 2nd derivative penalty is weighted linear in the observations $\{Y_i\}$. Silverman (1984) showed that the effective weight function looks like a kernel in a certain sense. The difficulty of spline smoothing is that the smoother is defined implicitly as the solution to a functional minimization problem. It is hard to obtain an explicit form of the weight function and to study the behaviour of the estimates as well as the effect of smoother on the data values. However extension to local constant fit with *kth* derivative penalty is easier: to apply the same analysis to local polynomial fit with *kth* derivative penalty, first, all estimates are weighted linear function of observations Y_i. The proof of this is given for local linear fit with 2nd derivative penalty and extension to *kth* derivative is easy.

In fact, if both g and g_1 minimize $S(g, g_1)$ under local linear fit with 2nd derivative penalty, then

$$S(\hat{g} + \alpha g, \hat{g}_1 + \beta g_1) \geq S(\hat{g}, \hat{g}_1)$$

for any $g \in C^2$, $g_1 \in C^2$ and $(\alpha, \beta) \in R^2$. This means that the real surface function

$$T(\alpha, \beta) = S(\hat{g} + \alpha g, \hat{g}_1 + \beta g_1)$$

has a local minimum at $\alpha = \beta = 0$, thus,

$$T'_\alpha(0,0) = T'_\beta(0,0) = 0.$$

That is, for $t \in [a,b]$

$$\sum_i [Y_i - \hat{g}(t_i) - (t - t_i)\hat{g}_1(t_i)]g(t_i) + \lambda \int \hat{g}''(t)\, g''(t)\, dt = 0$$

$$\sum_i [Y_i - \hat{g}(t_i) - (t - t_i)\hat{g}_1(t_i)](t - t_i)g_1(t_i) + \lambda_1 \int \hat{g}_1''(t)\, g_1''(t)\, dt = 0 \qquad (8.28)$$

Consider now two couple of splines $(\hat{g}^{[1]}, \hat{g}_1^{[1]})$ and $(\hat{g}^{[2]}, \hat{g}_1^{[2]})$ for the data $\{(t_i, Y_i^{[1]})\}_1^n$ and $\{(t_i, Y_i^{[2]})\}_1^n$.

From equations (8.28) one can see that $(\hat{g}^{[1]} + \hat{g}^{[2]}, \hat{g}_1^{[1]} + \hat{g}_1^{[2]})$ are splines for data $\{(t_i, Y_i^{[1]} + Y_i^{[2])}\}_1^n$. If the data vector $\{Y_i\}_1^n$ is written as a linear combination of vectors of n coordinates, then it is easily seen that both $\hat{g}$ and $\hat{g}_1$ are weighted linear functionals of observations $\{Y_i\}_1^n$.

Thus for given $t \in [a,b]$, there exist weights $G(t,t_i)$ such that

$$g(t) = \frac{1}{n}\sum_j y_j G(t, t - t_j) \qquad (8.29)$$

(same for other g_i, $\quad i = 1, 2, \cdots, m$).

However, the functional form of $\{G(t,t_j)\}_1^n$ is extremely complicated, it depends on the smoothing parameters $\lambda, \cdots, \lambda_m$ and design points. To advance the argument the following conditions A are necessary. Same conditions can be found in studying the spline smoother of local constant fit with 2nd derivative penalty (Eubank, 1988).

Condition A

1) All $g(.)$ are functions in the kth order Sobolev space

$$W_2^k[0,1] = \{g : g^{(j)} \ \ is \ \ absolutely \ \ continuous, \ \ j = 0, 1, \cdots, k-1 \ \ and \ \ g^{(j)} \in L_2[0,1]\}$$

2) All $g(.)$ satisfy the periodic boundary conditions

$$g^{(j)}(0) = g^{(j)}(1), \quad j = 0, 1, \cdots, k-1$$

3) The design points $t_0, t_1, \cdots, t_n$ are uniformly spaced over [0,1], i.e

$$t_j = \frac{j}{n}, \quad j = 0, 1, \cdots, n$$

4) Let $C_j = \int_0^1 t g_1(t) exp(-2\pi i j t)\, dt$, then $lim_{n\to\infty} \sum_s |C_{j+sn}| = 0$.

Condition B

1), 2), 3) and 4') with 4') $g_1^{(j)}(0) = g_1^{(j)}(1) = 0, j = 0, 1, \cdots, k-1$.

We are treating periodic smoothing splines from the conditions A and B, since they provide an important theoretical tool from past experience although seldom used in practice (Wahba, 1975, Rice and Rosenblatt, 1981, and Eubank, 1988). As Eubank said: "this allows for a simplified treatment which embodies many of the key features of asymptotic analysis for smoothing splines with minimum of technical details" (Section 6.3, Eubank, 1988). The periodic case also provides a springboad whose understanding makes what transpires in the general case seem intuitively plausible.

Remark: According to Section 3.4.2 of Eubank (1988), any Fourier coefficient of function satisfying the 1), 2) and 3) of Condition A can be assumed to decay

algebraically, so the C_j in 4) of Condition A satisfies: $\sum_s |C_{j+sn}| < \infty$, but $lim_{n\to\infty} \sum_s |C_{j+sn}| = 0$ is stronger than this. Note that $\frac{d}{dj} \int g_1(t) exp(-2\pi ijt) dt = -2\pi i \int t g_1(t) exp(-2\pi ijt) dt$ and the Fourier coefficient B_j of $g_1(t)$ is the inner product of $g_1(t)$ and the orthogonal basis, so the 4) of Condition A just results in the variable rate of B_n approaches to zero, when n is large enough, that is, $g_1(t)$ can be determined by finite number of orthogonal basis of R^n. Condition (4) is stronger than condition (4') so that the weight function $G(t,u)$ in the following Theorem 8.4.1 is simpler than that of Theorem 8.4.2. Also, the 4) of Condition A is for local linear fitting, and the general condition instead of 4) for local mth polynomial fitting is $lim_{n\to\infty} \sum_s |C^{\nu}_{j+sn}| = 0$, with $C^{\nu}_j = \int_0^1 t^{\nu} g_{\nu}(t) exp(-2\pi ijt) dt$, $\nu = 1, 2, ..., m$.

Theorem 8.4.1: Under condition A, the weight function in equation (8.29) satisfies

$$G(t,u) = \int_{-\infty}^{+\infty} \frac{1}{(a(t) + \lambda\, x^{2k})} exp(2\pi ix(t-u))\, dx,\ k = 1, 2, \cdots$$

where $a(t)$ is a function of t and the degree of polynomial m only.

$$a(t) = 1 \tag{8.30}$$

local constant fit

$$a(t) = 1 + \frac{\lambda}{\lambda_1} t^2 \tag{8.31}$$

local linear fit

$$a(t) = 1 + \frac{\lambda}{\lambda_1} t^2 + \frac{\lambda}{\lambda_2} (\frac{1}{2} t^2)^2 \tag{8.32}$$

local quadratic fit

$$\ddots$$

$$a(t) = 1 + \frac{\lambda}{\lambda_1} t^2 + \frac{\lambda}{\lambda_2} (\frac{1}{2} t^2)^2 + \cdots + \frac{\lambda}{\lambda_m} (\frac{1}{m!} t^2)^m. \tag{8.33}$$

local mth degree polynomial fit

Theorem 8.4.2: Under condition B, the weight function in equation (8.29) is given by

$$G(t,u) = \int_{-\infty}^{\infty} [\frac{1}{(a(t) + \lambda\, x^{2k} + \sum_1^{2m} a_j(t)\, x^{-2j})}]\, \exp(2\pi i x(t-u))\, dx \qquad (8.34)$$

where $a_j(t)$ $(j = 1, 2, \cdots, 2m)$ are complicated functions depending on the ratios $\frac{\lambda}{\lambda_1}, \cdots, \frac{\lambda}{\lambda_m}$ and t, $a_j(t)$ is real when j is even, otherwise imaginary number.

Proof: First consider local linear fit with *kth* derivative penalty under condition A. The smoothers in this case are minimizers of

$$\begin{aligned} &\frac{1}{n}\sum_1^n (y_j - g(t_j) - (t-t_j)g_1(t_j))^2 \\ &\quad + \frac{\lambda}{(-1)^k(2\pi)^{2k}} \int_0^1 g^{(2k)}(t)^2\, dt \\ &\quad + \frac{\lambda_1}{(-1)^k(2\pi)^{2k}} \int_0^1 g_1^{(2k)}(t)^2\, dt \end{aligned} \qquad (8.35)$$

over all g and $g_1 \in W_2^k[0,1]$. The factor of $(-1)^k(2\pi)^{2k}$ has been introduced into the criterion simply for subsequent notational convenience.

Using Fourier expansion of $g(t)$, then

$$g(t) = \sum_{-\infty}^{+\infty} A_j exp(2\pi ijt) \qquad (8.36)$$

and A_j is jth Fourier coefficient

$$A_j = \int_0^1 g(t) exp(-2\pi ijt) dt \qquad (8.37)$$

Similarly,

$$g_1(t) = \sum_{-\infty}^{+\infty} B_j exp(2\pi ijt) \qquad (8.38)$$

with

$$B_j = \int_0^1 g_1(t) exp(-2\pi ijt) dt \qquad (8.39)$$

and to find the minimizers of (8.35) it suffices to find its corresponding Fourier coefficients. The details of expressing (8.37) in term of Fourier coefficients of g and g_1 are given below for odd n only, similar approach is applied for even n.

When n is odd the vectors

$$\mathbf{X_j} = (1, exp(2\pi i\frac{j}{n}), \cdots, exp(2\pi i\frac{j(n-1)}{n}))$$

$$j = 0, \pm 1, \cdots, \pm\frac{(n-1)}{2}$$

form an orthogonal basis for R^n. Thus every vector in R^n can be expressed as a combination of vectors $\mathbf{X_j}$, particularly,

$$\beta_j = \frac{1}{n}\sum_1^n y_r exp(-2\pi ij\frac{(r-1)}{n}) \tag{8.40}$$

Then

$$y_k = \sum_{|j|\leq\frac{(n-1)}{2}} \beta_j exp(2\pi ij\frac{(k-1)}{n}) \tag{8.41}$$

$$g(t_k) = \sum_{|j|\leq\frac{(n-1)}{2}}\sum_{s=-\infty}^{+\infty} A_{j+sn} exp(2\pi ij\frac{(k-1)}{n}) \tag{8.42}$$

$$g_1(t_k) = \sum_{|j|\leq\frac{(n-1)}{2}}\sum_{s=-\infty}^{+\infty} B_{j+sn} exp(2\pi ij\frac{(k-1)}{n}) \tag{8.43}$$

$k = 1, 2, \cdots, n.$

Now, under condition A, we have

$$\begin{aligned}&\tfrac{1}{n}\sum_1^n[y_k - g(t_k) - (t-t_k)g_1(t_k)]^2\\ &= \sum_{|j|\leq\frac{(n-1)}{2}} |\beta_j - \sum_{s=-\infty}^{+\infty} A_{j+sn} - t\sum_{s=\infty}^{+\infty} B_{j+sn} + \sum_{s=-\infty}^{+\infty} C_{j+sn}|^2.\end{aligned} \tag{8.44}$$

Also, $g^{(k)}(t)$ has Fourier coefficients

$$\int_0^1 g^{(k)}(t)exp(-2\pi ijt)dt = (2\pi ij)^k A_j \tag{8.45}$$

Therefore, by Parseval's relation

$$\begin{aligned}\int_0^1 g^{(k)}(t)^2 dt &= (-1)^k (2\pi)^{2k} \sum_{-\infty}^{+\infty} j^{2k} |A_j|^2 \\ &= (-1)^k (2\pi)^{2k} \sum_{|j| \le \frac{(n-1)}{2}} \sum_{s=-\infty}^{+\infty} (j+sn)^{2k} |A_{j+sn}|^2 \end{aligned} \tag{8.46}$$

Similarly,

$$\int_0^1 g_1^{(k)}(t)^2 dt = (-1)^k (2\pi)^{2k} \sum_{|j| \le \frac{(n-1)}{2}} \sum_{s=-\infty}^{+\infty} (j+sn)^{2k} |B_{j+sn}|^2 \tag{8.47}$$

Combining (8.44), (8.46) and (8.47) and under condition A, (8.35) can be approximately written as

$$\sum_{|j| \le \frac{(n-1)}{2}} (|\beta_j - \sum_s A_{j+sn} - t \sum_s B_{j+sn}|^2 + \lambda \sum_s (j+sn)^{2k} |A_{j+sn}|^2 + \lambda_1 \sum_s (j+sn)^{2k} |B_{j+sn}|^2) \tag{8.48}$$

which depends only on the Fourier coefficients and is a sum of n nonegative functions. For general j, the minimizers of

$$|\beta_j - \sum_s A_{j+sn} - t \sum_s B_{j+sn}|^2 + \lambda \sum_s (j+sn)^{2k} |A_{j+sn}|^2 + \lambda_1 \sum_s (j+sn)^{2k} |B_{j+sn}|^2 \tag{8.49}$$

are obtained by differentiation of this expression with respect to A_{j+sn} and B_{j+sn}. The optimal coefficients satisfy

$$\begin{aligned}(\beta_j - \sum_s A_{j+sn} - t \sum_s B_{j+sn}) &= \lambda (j+sn)^{2k} A_{j+sn} \\ (\beta_j - \sum_s A_{j+sn} - t \sum_s B_{j+sn}) t &= \lambda_1 (j+sn)^{2k} B_{j+sn} \end{aligned} \tag{8.50}$$

Thus,

$$B_{j+sn} = \frac{\lambda}{\lambda_1} t A_{j+sn}. \tag{8.51}$$

Setting

$$a(t) = 1 + \frac{\lambda}{\lambda_1} t^2$$

and substituting (8.51) into (8.50), then

$$A_{j+sn} = \frac{1}{\lambda (j+sn)^{2k}} (\beta_j - a(t) \sum_s A_{j+sn}) \tag{8.52}$$

Summing this expression over s to obtain

$$\sum_{s=-\infty}^{+\infty} A_{j+sn} = \beta_j \frac{r_j(k)}{\lambda + a(t) r_j(k)} \tag{8.53}$$

where

$$r_j(k) = \sum_s (j+sn)^{-2k}$$

Substituting back into the original formula for A_{j+sn} then the minimizing coefficients are

$$A_{j+sn} = \frac{\beta_j}{(\lambda + a(t) r_j(k))} \frac{1}{(j+sn)^{2k}} \tag{8.54}$$

Now the smoothing spline estimator is given explicitly, for n odd, as

$$\begin{aligned} g(t) &= \sum_{|j| \le \frac{(n-1)}{2}} \sum_{s=-\infty}^{+\infty} A_{j+sn} exp(2\pi i (j+sn) t) \\ &= \sum_{|j| \le \frac{(n-1)}{2}} \frac{\beta_j}{(\lambda + a(t) r_j(k)) j^{2k}} exp(2\pi i j t) \\ &+ \sum_{|j| \le \frac{(n-1)}{2}} \left(\frac{\beta_j}{(\lambda + a(t) r_j(k))} \right) \sum_{s \neq 0} (j+sn)^{-2k} exp(2\pi i (j+sn) t) \end{aligned} \tag{8.55}$$

Note that

$$|\sum_{s \neq 0} (j+sn)^{-2k} exp(2\pi i (j+sn))| \le \sum_{s \neq 0} (j+sn)^{-2k} = O(n^{-2k})$$

and

$$(\lambda + a(t) r_j(k)) j^{2k} = a(t) + \lambda j^{2k} + a(t) j^{2k} O(n^{-2k})$$

These results coupled with the fact that the Fourier coefficients for $g \in W_2^k[0,1]$ decay rapidly to zero suggest that

$$g(t) \approx \sum_{|j| \le \frac{(n-1)}{2}} \frac{\beta_j}{(a(t) + \lambda j^{2k})} exp(2\pi i j t) \tag{8.56}$$

for sufficiently large n. Further, substituting the (8.40) into (8.56), we have

$$g(t) \approx \frac{1}{n} \sum_{r=1}^{n} y_r \sum_{|j| \le \frac{(n-1)}{2}} (a(t) + \lambda j^{2k})^{-1} exp(2\pi i j (t - \frac{(r-1)}{n})) \tag{8.57}$$

and, for n large,

$$\begin{aligned} G(t,u) &= \sum_{|j|\leq\frac{(n-1)}{2}} (a(t)+\lambda j^{2k})^{-1} exp(2\pi ij(t-u)) \\ &\approx \int_{-\infty}^{+\infty} (a(t)+\lambda x^{2k})^{-1} exp(2\pi ix(t-u))dx \end{aligned} \tag{8.58}$$

Now, turn our discussion into that under condition B.

If let D_j be the jth Fourier coefficient of $tg_1''(t)$, then from the 4') of Condition B, we have $D_j = 2\pi ij(-2B_j + 2\pi ijC_j)$, so that

$$\sum_s \frac{D_{j+sn}}{4\pi^2(j+sn)^2} = -\sum_s C_{j+sn} + \sum_s \frac{B_{j+sn}}{\pi i(j+sn)}.$$

Then, from

$$D_0 = \int_0^1 tg_1''(t)dt = 0$$

and

$$lim_{n\to\infty} \sum_s |\frac{D_{j+sn}}{4\pi^2(j+sn)^2}| = 0,$$

we have

$$\sum_s C_{j+sn} \approx -\sum_s \frac{B_{j+sn}i}{\pi(j+sn)}.$$

Using similar calculations to (8.48), result in expressing (8.35) as

$$\begin{aligned} &\Sigma_{|j|\leq\frac{(n-1)}{2}} |\beta_j - \sum_s A_{j+sn} - t\sum_s B_{j+sn} - \frac{1}{\pi}i\sum_s \frac{B_{j+sn}}{j+sn}|^2 \\ &+ \sum_{|j|\leq\frac{(n-1)}{2}} \sum_s (j+sn)^{2k}[\lambda|A_{j+sn}|^2 + \lambda_1|B_{j+sn}|^2] \end{aligned} \tag{8.59}$$

The optimal coefficients A_{j+sn} and B_{j+sn} which minimize above equation are obtained by differentiation and for $j \geq 1$ these satisfy

$$(\beta_j - \sum_s A_{j+sn} - t\sum_s B_{j+sn} - \frac{i}{\pi}\sum_s \frac{B_{j+sn}}{j+sn}) = \lambda(j+sn)^{2k}A_{j+sn} \tag{8.60}$$

and

$$(t+\frac{i}{\pi(j+sn)})(\beta_j - \sum_s A_{j+sn} - t\sum_s B_{j+sn} - \frac{i}{\pi}\sum_s \frac{B_{j+sn}}{j+sn}) = \lambda_1(j+sn)^{2k}B_{j+sn} \tag{8.61}$$

Thus

$$B_{j+sn} = \frac{\lambda}{\lambda_1}\Big(t + \frac{i}{\pi(j+sn)}\Big)A_{j+sn} \tag{8.62}$$

and

$$A_{j+sn} = \frac{1}{\lambda(j+sn)^{2k}}\Big[\beta_j - a(t)\sum_r A_{j+rn} - ib(t)\sum_r \frac{A_{j+rn}}{j+rn} + c(t)\sum_r \frac{A_{j+rn}}{(j+rn)^2}\Big] \tag{8.63}$$

where

$$b(t) = \frac{2}{\pi}\frac{\lambda}{\lambda_1}t$$

and

$$c(t) = \frac{1}{\pi^2}\frac{\lambda}{\lambda_1}.$$

Further calculation of A_{j+sn} , $\sum_r \frac{A_{j+rn}}{j+rn}$ and $\sum_r \frac{A_{j+rn}}{(j+rn)^2}$ leads to

$$\begin{aligned} g(t) \;&=\; \sum_{1\le|j|\le\frac{(n-1)}{2}} \sum_s \frac{\beta_j}{(j+sn)^{2k}(\lambda + a(t)r_j(k) + ib(t)r_j(k+1) - c(t)r_j(k+2))} \\ &\quad .\; exp(2\pi i(j+sn)t). \end{aligned} \tag{8.64}$$

Note that

$$\sum_s \frac{1}{(j+sn)^{2k}(\lambda + a(t)r_j(k) + ib(t)r_j(k+1) - c(t)r_j(k+2))}$$

$$= \frac{r_j(k)}{\lambda + a(t)r_j(k) + ib(t)r_j(k+1) - c(t)r_j(k+2)}.$$

Writing

$$r_j(k) = j^{-2k}(1 + \sum_{s\neq 0}(\frac{j}{j+sn})^{2k},$$

then

$$\sum_{1\le|j|\le\frac{(n-1)}{2}} r_j(k+1)/r_j(k) = \sum_{1\le|j|\le\frac{(n-1)}{2}} \frac{1}{j^2}\frac{1+\sum_{s\neq 0}(\frac{j}{j+sn})^{2k}}{1+\sum_{s\neq 0}(\frac{j}{j+sn})^{2k+2}},$$

and approximately

$$\sum_{1\le|j|\le\frac{(n-1)}{2}} r_j(k+1)/r_j(k) = \int x^{-2}dx,$$

then approximately argue the weighting of $\{Y_i\}$ in equation (8.64) as

$$(\int_{-\infty}^{+\infty})(\frac{1}{a(t)+\lambda\, x^{2k}+ib(t)\, x^{-2}-c(t)\, x^{-4}})exp(2\pi ix(t-u_i))dx \qquad (8.65)$$

This completes the proof under local linear fit with kth derivative penalty.

To prove the above Theorems (8.4.1) and (8.4.2) for local quadratic fit with kth derivative penalty, the smoothers minimize the following expression

$$\begin{aligned}\frac{1}{n}\ & \Sigma_1^n\ (y_j-g(t_j)-(t-t_j)g_1(t_j)-\frac{1}{2}(t-t_j)^2g_2(t_j))^2 \\ & + \frac{\lambda}{(-1)^k(2\pi)^{2k}}\int_0^1 g^{(2k)}(t)^2\ dt \\ & + \frac{\lambda_1}{(-1)^k(2\pi)^{2k}}\int_0^1 g_1^{(2k)}(t)^2\ dt \\ & + \frac{\lambda_2}{(-1)^k(2\pi)^{2k}}\int_0^1 g_2^{(2k)}(t)^2\ dt \end{aligned} \qquad (8.66)$$

and these theorems can be proved along the same lines as before as long as we transfer the conditions on $g_1(t)$ into $g_2(t)$. However, only Theorem 8.4.2 is proved in the following as it requires more details.

Proof of Theorem 8.4.2 (local quadratic fit): Consider

$$\frac{1}{2}(t-t_j)^2g_2(t_j)=\frac{1}{2}t^2\ g_2(t_j)-t\ t_jg_2(t_j)+\frac{1}{2}\ t_j^2g_2(t_j) \qquad (8.67)$$

and let E_j be the jth Fourier coefficient of $g_2(t)$, and from

$$\begin{aligned}\int_0^1 t^2g_2(t)exp(-2\pi ijt)dt &= -\frac{i}{2\pi j}E_j+\frac{1}{2\pi ij}\int_0^1 tg_2'(t)exp(-2\pi ijt)dt \\ \int_0^1 tg_2(t)exp(-2\pi ijt)dt &= -\frac{i}{\pi j}\int_0^1 t^2g_2(t)exp(-2\pi ijt)dt \\ &+ \frac{1}{2\pi ij}\int_0^1 t^2g_2'(t)exp(-2\pi ijt)dt,\end{aligned}$$

and using same argument as (8.59), then (8.66) can be written as

$$\Sigma_{1\le|j|\le\frac{(n-1)}{2}}\ |\beta_j-\sum_s A_{j+sn}-t\sum_s B_{j+sn}-\frac{1}{2}t^2\sum_s E_{j+sn}-\frac{1}{\pi}i\sum_s\frac{B_{j+sn}}{j+sn}$$

$$
\begin{aligned}
- \quad & \frac{1}{\pi} ti \sum_s \frac{E_{j+sn}}{j+sn} + \frac{1}{4\pi^2} \sum_s \frac{E_{j+sn}}{(j+sn)^2}|^2 \\
+ \quad & \sum_{|j| \leq \frac{(n-1)}{2}} \sum_s (j+sn)^{2k} [\lambda |A_{j+sn}|^2 + \lambda_1 |B_{j+sn}|^2 + \lambda_2 |E_{j+sn}|^2] \qquad (8.68)
\end{aligned}
$$

The differentiation of this expression with respect to A_{j+sn}, B_{j+sn} and E_{j+sn}, resulting in the minimize coefficient of (8.68) satisfy

$$
\begin{aligned}
& \Big(\beta_j - \sum_s A_{j+sn} - t \sum_s B_{j+sn} - \frac{1}{2} t^2 \sum_s E_{j+sn} \\
& - \frac{i}{\pi} \sum_s \frac{B_{j+sn}}{j+sn} - \frac{i}{2\pi} t \sum_s \frac{E_{j+sn}}{j+sn} + \frac{1}{4\pi^2} \sum_s \frac{E_{j+sn}}{(j+sn)^2} \Big) \\
= \quad & \lambda (j+sn)^{2k} A_{j+sn} \\
& \Big(\beta_j - \sum_s A_{j+sn} - t \sum_s B_{j+sn} - \frac{1}{2} t^2 \sum_s E_{j+sn} - \frac{i}{\pi} \sum_s \frac{B_{j+sn}}{j+sn} \\
& - \frac{i}{2\pi} t \sum_s \frac{E_{j+sn}}{j+sn} + \frac{1}{4\pi^2} \sum_s \frac{E_{j+sn}}{(j+sn)^2} \Big) \Big(t + \frac{i}{\pi(j+sn)} \Big) \\
= \quad & \lambda_1 (j+sn)^{2k} B_{j+sn} \\
& \Big(\beta_j - \sum_s A_{j+sn} - t \sum_s B_{j+sn} - \frac{1}{2} t^2 \sum_s E_{j+sn} \\
& - \frac{i}{2\pi} \sum_s \frac{B_{j+sn}}{j+sn} - \frac{i}{2\pi} t \sum_s \frac{E_{j+sn}}{(j+sn)} + \frac{1}{4\pi^2} \sum_s \frac{E_{j+sn}}{(j+sn)^2} \Big) \\
& \Big(\frac{1}{2} t^2 + \frac{i}{2\pi(j+sn)} t - \frac{1}{4\pi^2 (j+sn)^2} \Big) \\
= \quad & \lambda_2 (j+sn)^{2k} E_{j+sn}
\end{aligned}
$$

and

$$
\begin{aligned}
B_{j+sn} &= \frac{\lambda}{\lambda_1} \Big(t + \frac{i}{\pi(j+sn)} \Big) A_{j+sn} \\
E_{j+sn} &= \frac{\lambda}{\lambda_2} \Big(\frac{1}{2} t^2 + \frac{i}{2\pi(j+sn)} t - \frac{1}{4\pi^2 (j+sn)^2} \Big) A_{j+sn}
\end{aligned}
$$

Then

$$
\begin{aligned}
A_{j+sn} = \frac{1}{\lambda (j+sn)^{2k}} \Big[& \beta_j - a(t) \sum_s A_{j+sn} - a_1(t) \sum_s \frac{A_{j+sn}}{(j+sn)} \\
- \quad & a_2(t) \sum_s \frac{A_{j+sn}}{(j+sn)^2} - a_3(t) \sum_s \frac{A_{j+sn}}{(j+sn)^3} - a_4(t) \sum_s \frac{A_{j+sn}}{(j+sn)^4} \Big] \qquad (8.69)
\end{aligned}
$$

where

$$
\begin{aligned}
a_1(t) &= \frac{\lambda}{\pi}(\frac{1}{\lambda_1} + \frac{1}{\lambda_2}\frac{1}{2}t^2)\ ti \\
a_2(t) &= -\frac{\lambda}{2\pi^2}(\frac{1}{2\lambda_1} + \frac{1}{2\lambda_2}t^2 + \frac{1}{2\lambda_2}t^2) \\
a_3(t) &= -\frac{\lambda}{\lambda_2}\frac{1}{2\pi^3}t\ i \\
a_4(t) &= \frac{\lambda}{\lambda_2}\frac{1}{16\pi^4}
\end{aligned}
$$

Last, we have

$$
\begin{aligned}
g(t) = & \sum_{1\leq|j|\leq\frac{(n-1)}{2}} \sum_s \frac{\beta_j}{(j+sn)^{2k}} \\
\times & \frac{1}{(\lambda + a(t)r_j(k) + a_1(t)r_j(k+1) + a_2(t)r_j(k+2) + a_3(t)r_j(k+3) + a_4(t)r_j(k+4))} \\
\times & \ exp(2\pi i(j+sn)t)
\end{aligned}
$$

and Theorem 8.4.2 is obvious.

For general *mth* polynomial fit with derivative penalty, an expression for A_{j+sn} could be obtained along the same lines as above.

8.5 Equivalent Kernel and the Boundary Properties of the Method

It is known that the spline smoothing under local constant fit with 2nd derivative penalty corresponds approximately to smoothing by a kernel method (Silverman, 1984), and bandwidth and order of the kernel are $\lambda^{\frac{1}{4}}$ and 4 respectively. First extension to local constant fit with derivative penalty is carried out by generalizing to any degree local polynomial fit with any derivative penalty. As in Section 8.4

and under local constant fit with kth derivative penalty, the weight function is approximated under condition A by

$$G(t,u) = \lambda^{-\frac{1}{2k}} K_0(\frac{t-u}{\lambda^{\frac{1}{2k}}}) \tag{8.70}$$

with

$$K_0(z) = \int_{-\infty}^{+\infty} (\frac{1}{1+x^{2k}})\, exp(2\pi i x z)\, dx \tag{8.71}$$

and the function $K_0(u)$ can be regarded as the Fourier transformation of $f_0(x)$

$$f_0(x) = \frac{1}{1+x^{2k}}. \tag{8.72}$$

Thus, in the interior points, via the derivatives of the Fourier transformation evaluated at zero, the bandwidth and order of equivalent kernel are $\lambda^{\frac{1}{2k}}$ and $2k$, respectively. This also is in Silverman (1984).

Now for general *mth* polynomial fit with *kth* derivative penalty, the approximate weight function under condition A is given by weighting as

$$\begin{aligned} G(t,u) &= \int_{-\infty}^{+\infty} (\frac{1}{a(t)+\lambda x^{2k}})\, exp(2\pi i x(t-u))\, dx \\ &= \frac{1}{a(t)\left(\frac{\lambda}{a(t)}\right)^{1/2k}} K_0\left(\frac{t-u}{\left(\frac{\lambda}{a(t)}\right)^{1/2k}}\right) \end{aligned} \tag{8.73}$$

For example,

$$k=1 \quad K_0(z) = \frac{1}{2}e^{-|u|};$$

$$k=2 \quad K_0(z) = \frac{1}{2}e^{-|u|/\sqrt{2}})sin(|u|/\sqrt{2}+\pi/4);$$

$$k=3 \quad K_0(z) = \frac{1}{6}\left(e^{-|u|} + 2e^{-|u|/2}sin(\sqrt{3}|u|/2+\pi/6)\right).$$

Figure 8.1 displays these three kernels and they are very similar in appearance.

Similarly, approximate weight function under condition B with *mth* polynomial fit of *kth* derivative penalty is

$$\frac{1}{a(t)\left(\frac{\lambda}{a(t)}\right)^{1/2k}} K_m\left(\frac{t-u}{\left(\frac{\lambda}{a(t)}\right)^{1/2k}}\right) \tag{8.74}$$

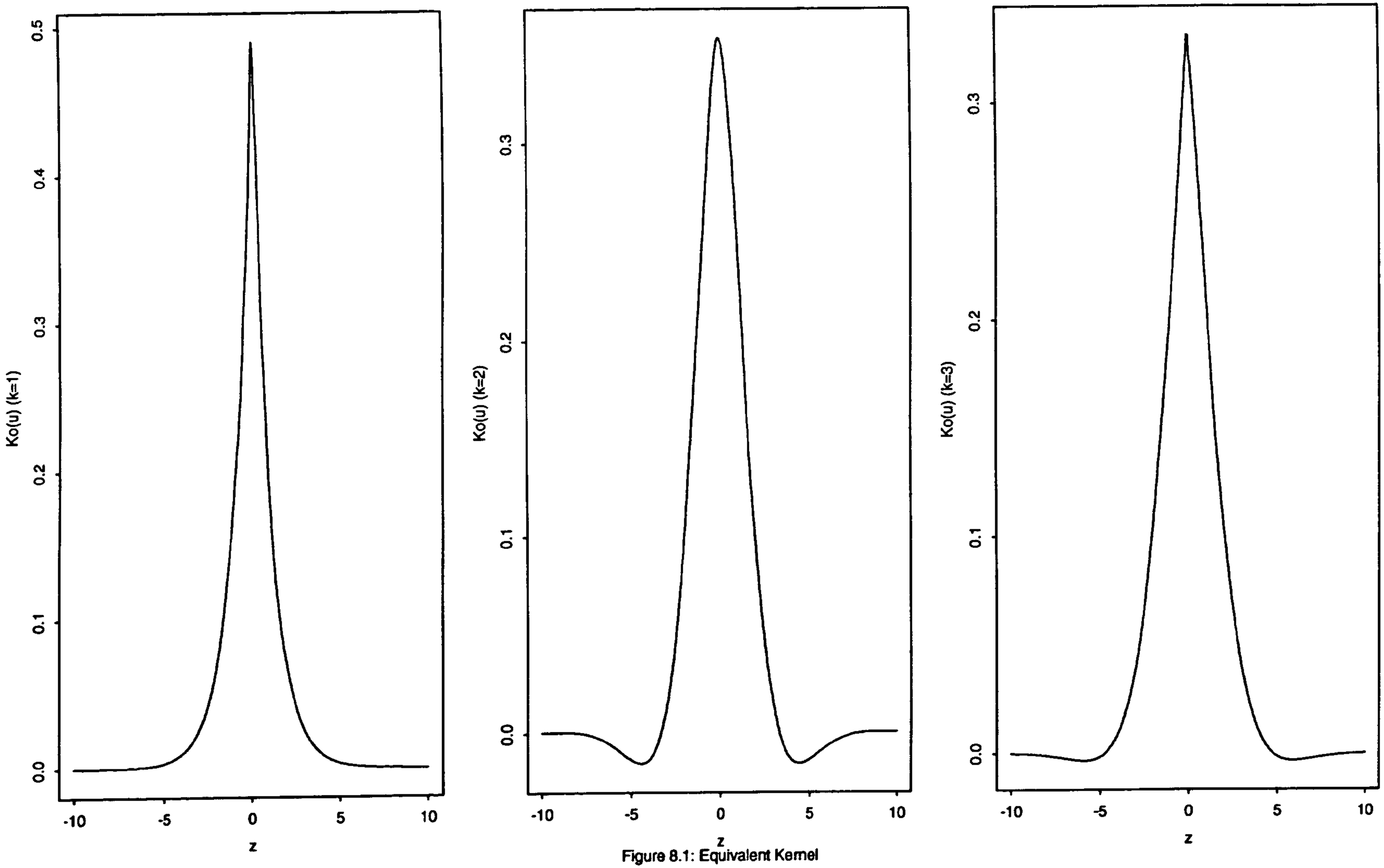


Figure 8.1: Equivalent Kernel

where

$$K_m(z) = \int_{-\infty}^{+\infty} f(z)\ exp(2\pi i x z)\ dx \tag{8.75}$$

and

$$f(z) = \frac{1}{1 + z^{2k} + \sum_1^{2m} a_j(t) \left(\frac{\lambda}{a(t)}\right)^{\frac{j}{2k}} z^{-j}} \tag{8.76}$$

Theorem 8.5.1: Under the uniform design, the bandwidth and the order of equivalent kernel for mth polynomial fit with kth derivative penality are $\left(\frac{\lambda}{a(t)}\right)^{1/2k}$ and $2k$, respectively.

When $m = 0$ (local constant fit), the bandwidth is $\lambda^{1/2k}$, which is independent of t from Theorem 8.5.1 so that the weight function $G(t, u)$ lacks the ability to change bandwidth particularly at boundary points. As Silverman (1984) said that weight function $G(t, u)$ deteriorates near the boundary of the design set ($u = 0$ or 1). These drawbacks, however, can be overcomed by using local polynomial fit. In fact, at a particular point, the bandwidth $\left(\frac{\lambda}{a(t)}\right)^{1/2k}$ is always related to t, the asymptotic weights, equivalent kernel, and bias and variance of the estimator near the boundary of support could be obtained by setting the point to $t = 0$ and $t = 1$. We can regard this bandwidth as a kind of variable bandwidth. It was shown that kernel estimators with such a local bandwidth choice are superior to the ordinary kernel estimators with global bandwidth choice if optimal bandwidths are used (Müller & Stadtmüller, 1987).

When $t = 0$, under the condition B, the approximat weight function (Theorem 8.4.2)

$$G(0, u) = \int_{-\infty}^{+\infty}) \frac{1}{1 + x^{2k} + \sum_1^m a_{2j}(0) x^{-2j}}\ exp(2\pi i x(-u))\ dx, \tag{8.77}$$

and all the $a_{2j}(0)$ $(j = 1, 2, \cdots, m)$ are real numbers, since all the imaginary numbers $a_{2j-1} = 0$ $(j = 1, 2, \cdots, m)$ in the expression (8.34) when $t = 0$. Thus, a

simpler expression for the weight and kernel is obtained at the boundaries, and the bandwidth and order of equivalent kernel in boundary are the same as those for interior points. Also, the algorithm for boundary splines is still easy by setting $t = 0$ and $t = 1$ in the diagmatrix W of Section 8.3. In the next section, it will be shown further that the order for bias in boundary is same as in interior by using mth $(m \geq 1)$ polynomial fit.

Generally, according to the hint of Silverman's work (1984), with design density $h(t)$ instead of uniform density, we should have the approximate weight function is

$$\frac{1}{h(t)a(t)\left(\frac{\lambda}{h(t)a(t)}\right)^{1/2k}} K_0\left(\frac{t-u}{\left(\frac{\lambda}{h(t)a(t)}\right)^{1/2k}}\right)$$

with bandwidth $\left(\frac{\lambda}{h(t)a(t)}\right)^{1/2k}$ (But we don't prove it at this moment).

In order to illustrate how well the approximation works out in practice, for example, under Condition A, some explicit calculations for weight function are carried out. A hundred design points t_i are place uniformly in interval [0, 1], and let $h(t)$ be $N(0.5, 0.25^2)$. The weight function $G(t, u)$ for 3 values: $i = 50, t_{50} = 0.464$, $i = 100, t_{100} = 0.967$ and $i = 1, t_1 = 0.06$ are displayed respectively in Figures 8.2, 8.3 and 8.3 with (a), (b), (c), which respectively correspond to local constant fitting, local linear fitting and local quadratic fitting, and solid curves always denote the exact weight function which can be got through smoothing data $\{t_l, y_l\}_{l=1}^n$ with all $y_l = 0$ except $y_i = n$. The values $\lambda = 10^{-7}$, $\lambda_1 = 2 \times 10^{-7}$ and $\lambda_2 = 4 \times 10^{-7}$ are used for smoothing parameters, as $\lambda_2 \geq \lambda_1 \geq \lambda$ is required generally in practice from the experience of kernel smoothing. From Figures 8.2, 8.3 and 8.4, it can be seen that the closeness of the approximation is remarkable for points t_i away from the boundary. For boundaries, the approximation is less good. However, three approximations (local constant fitting, local linear fitting

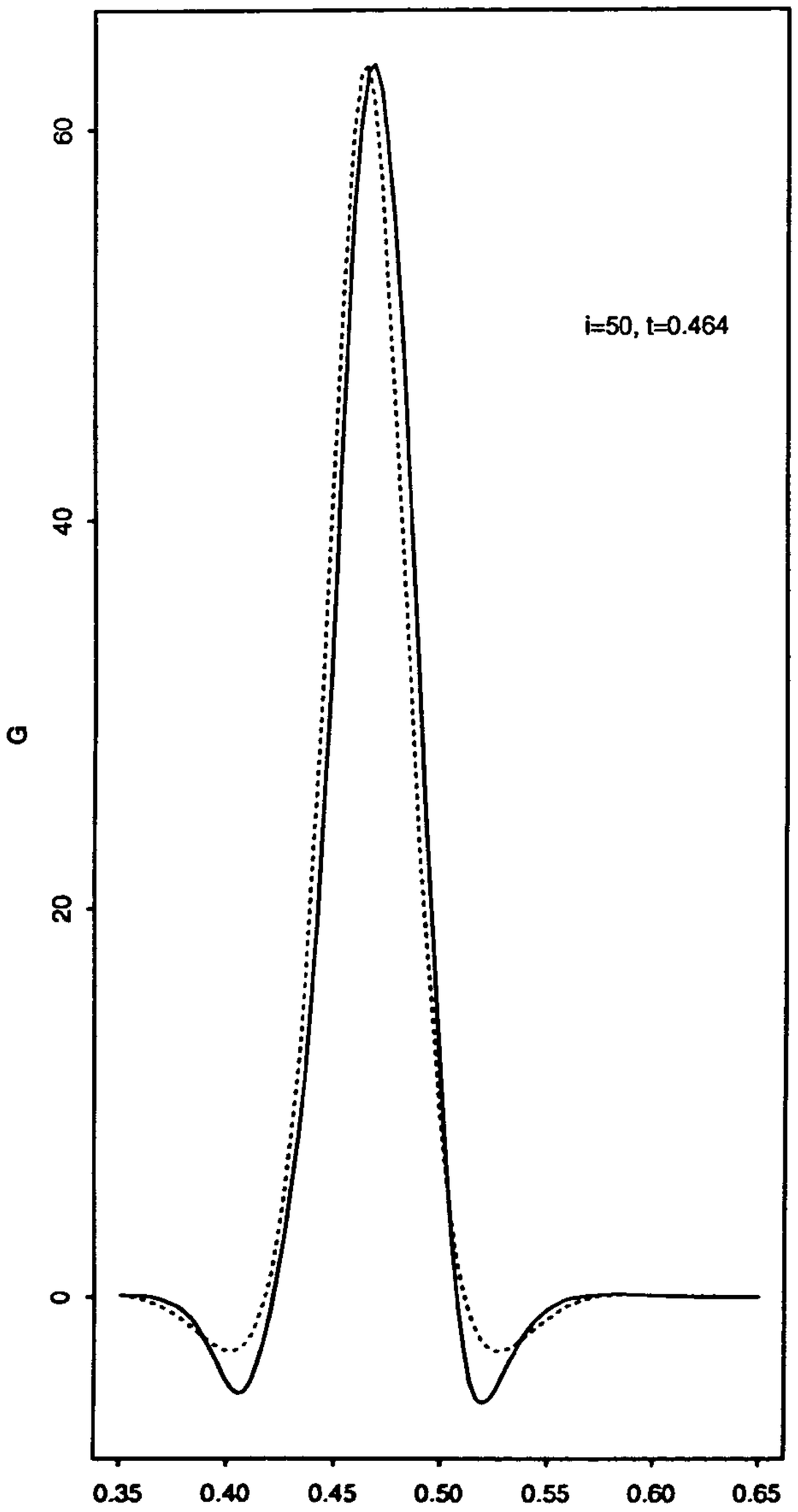


Figure 8.2 (a): Local constant fitting

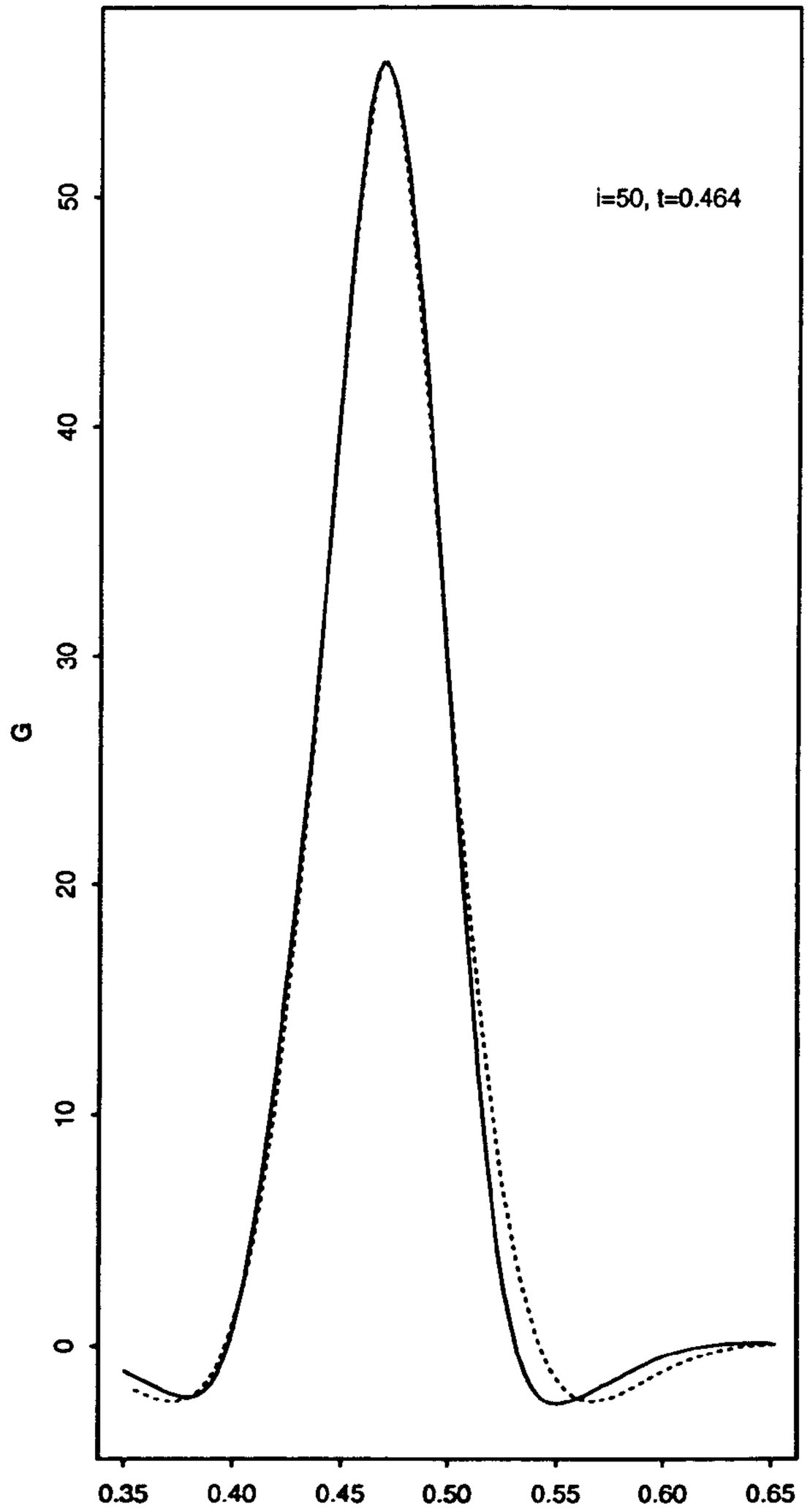


Figure 8.2 (b): Local linear fitting

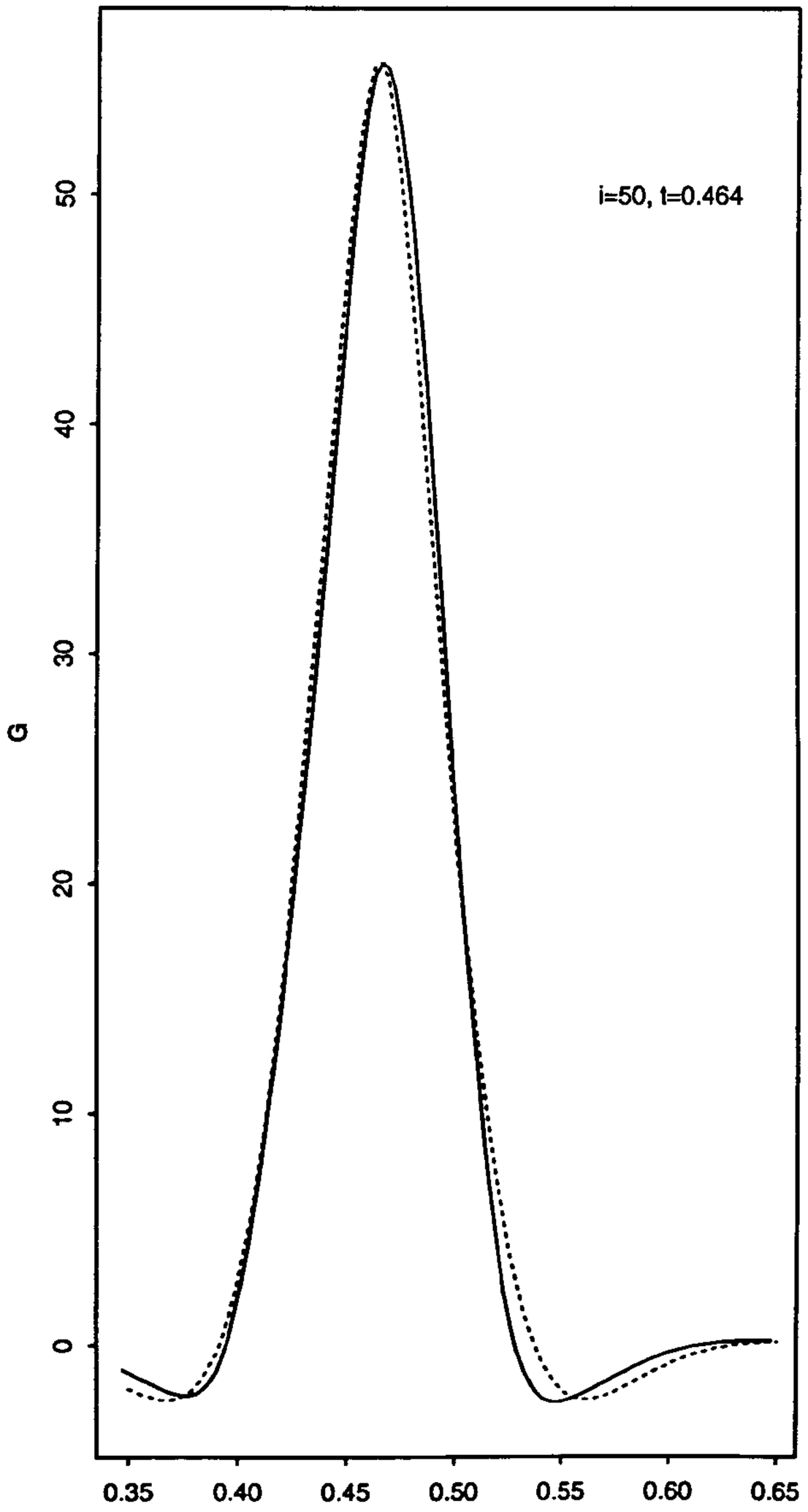


Figure 8.2 (b): Local quadratic fitting

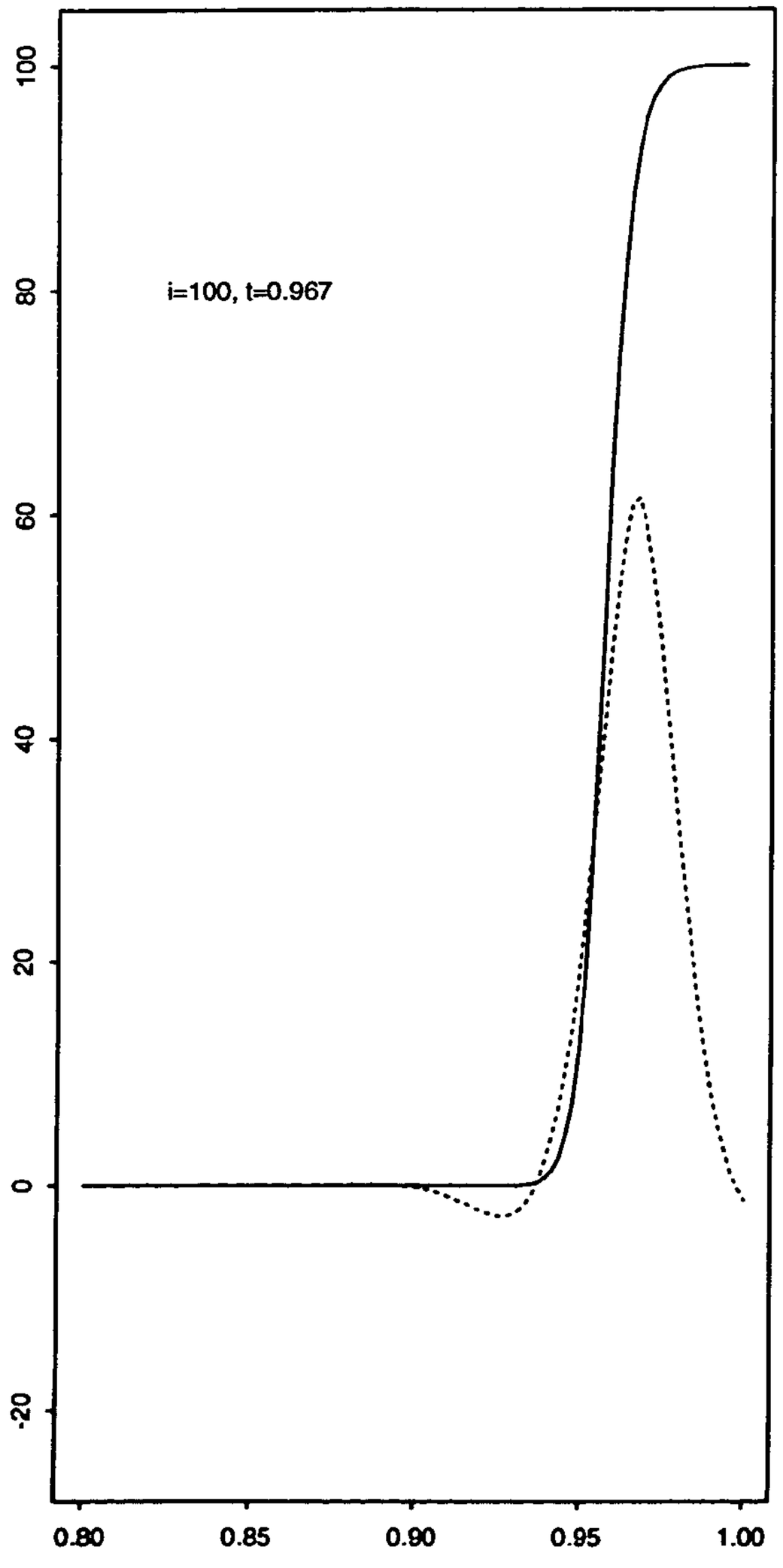


Figure 8.3 (a): Local constant fitting

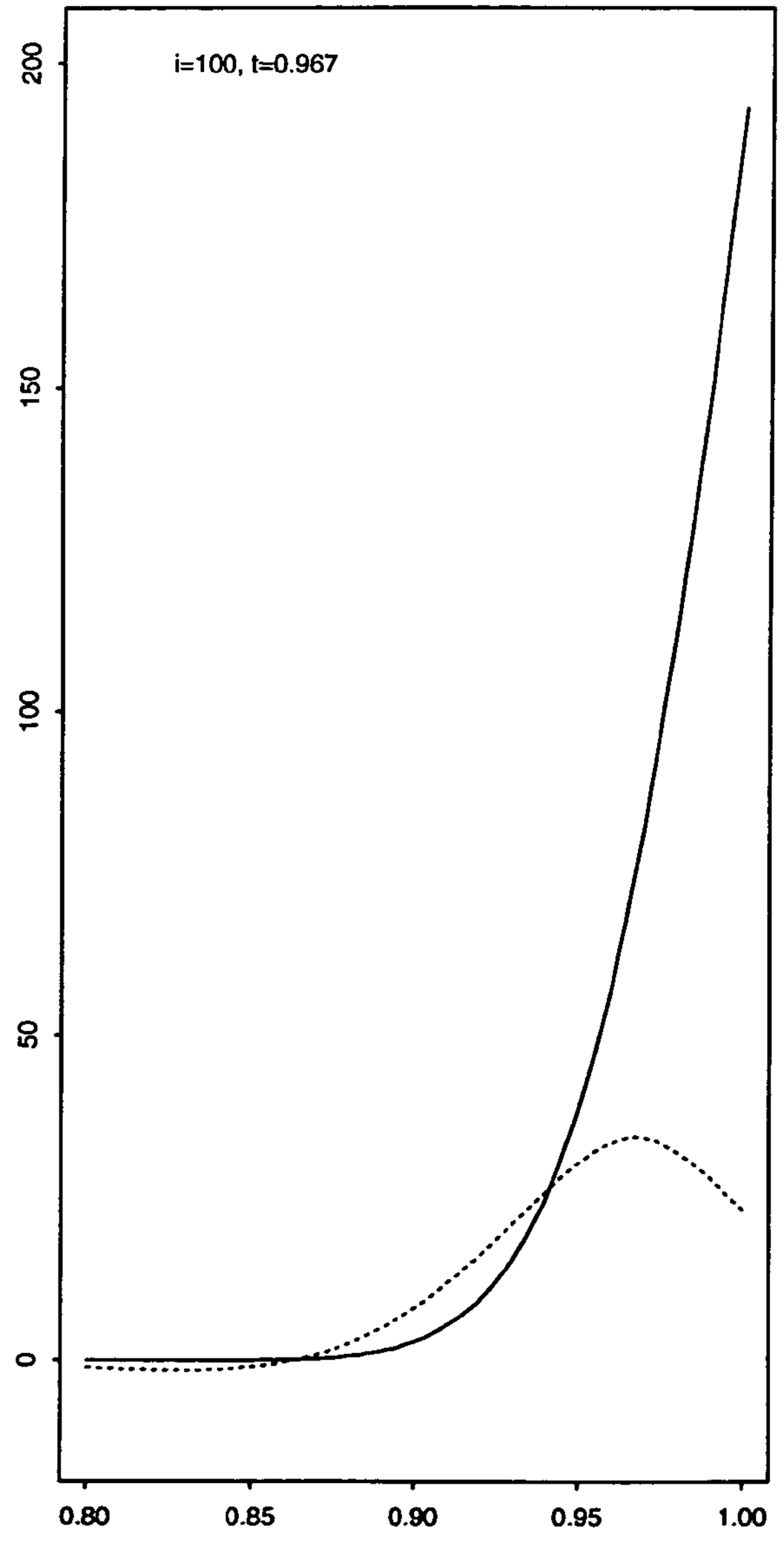


Figure 8.3 (b): Local linear fitting

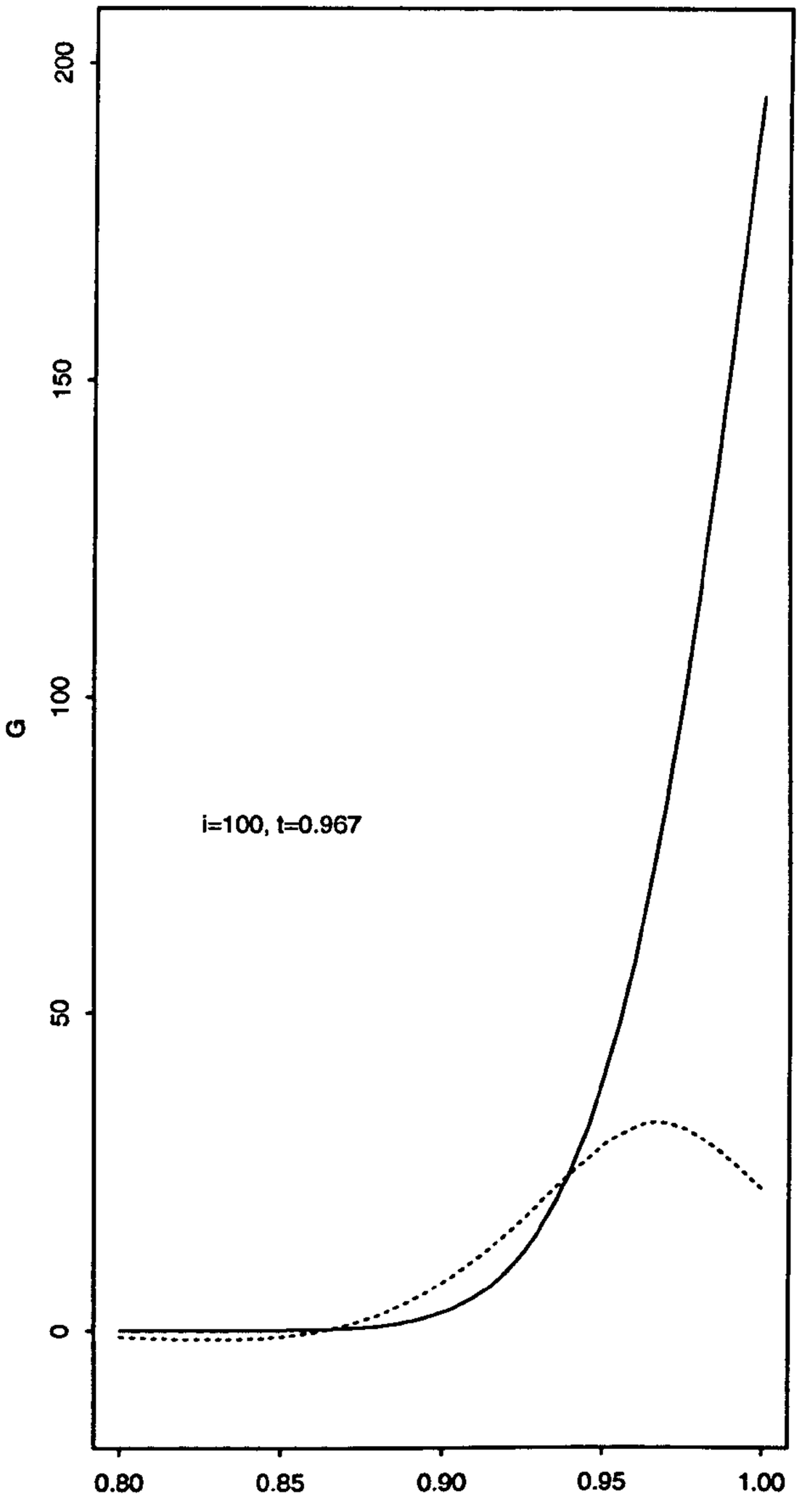


Figure 8.3 (c): Local quadratic fitting

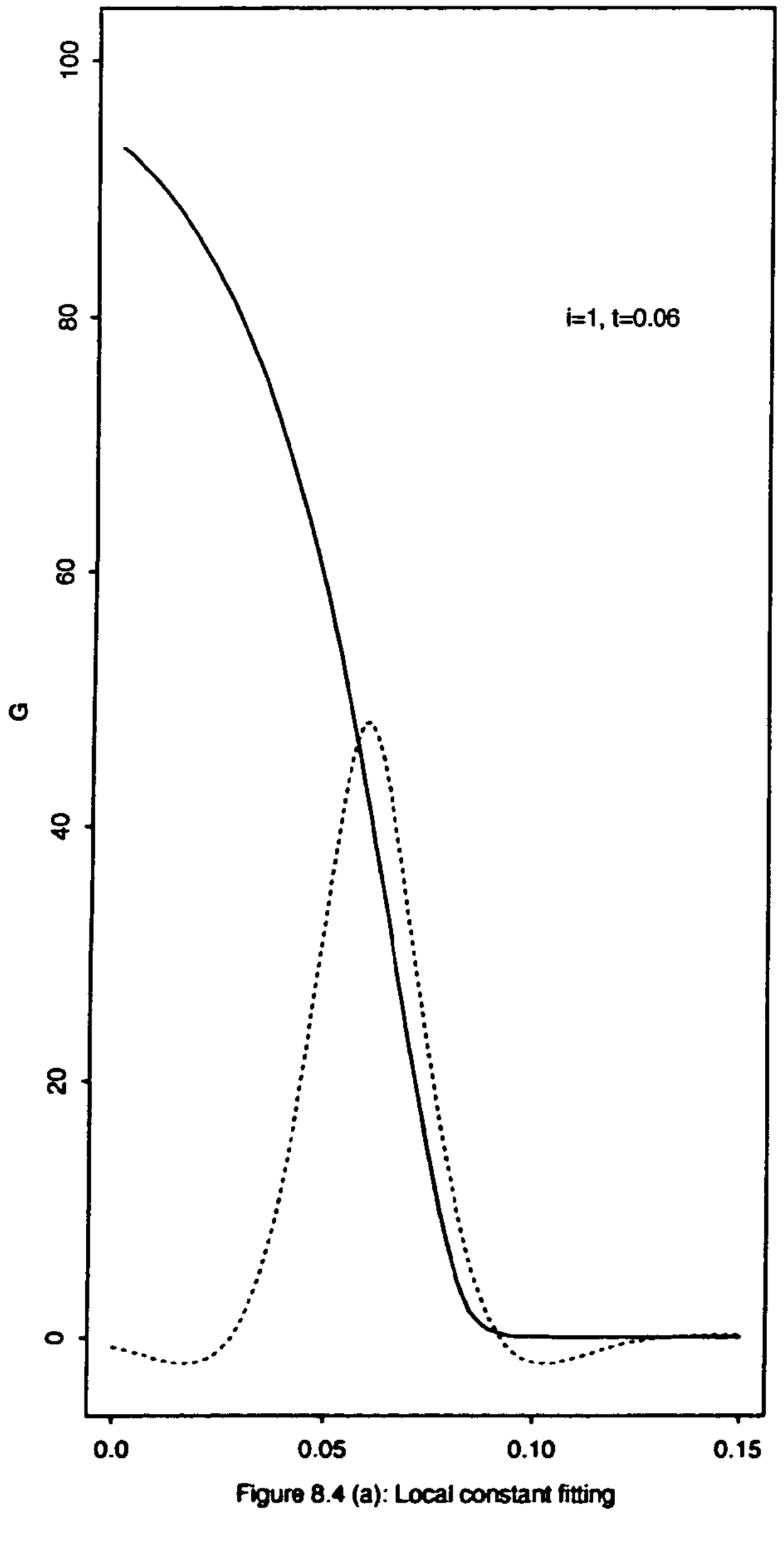


Figure 8.4 (a): Local constant fitting

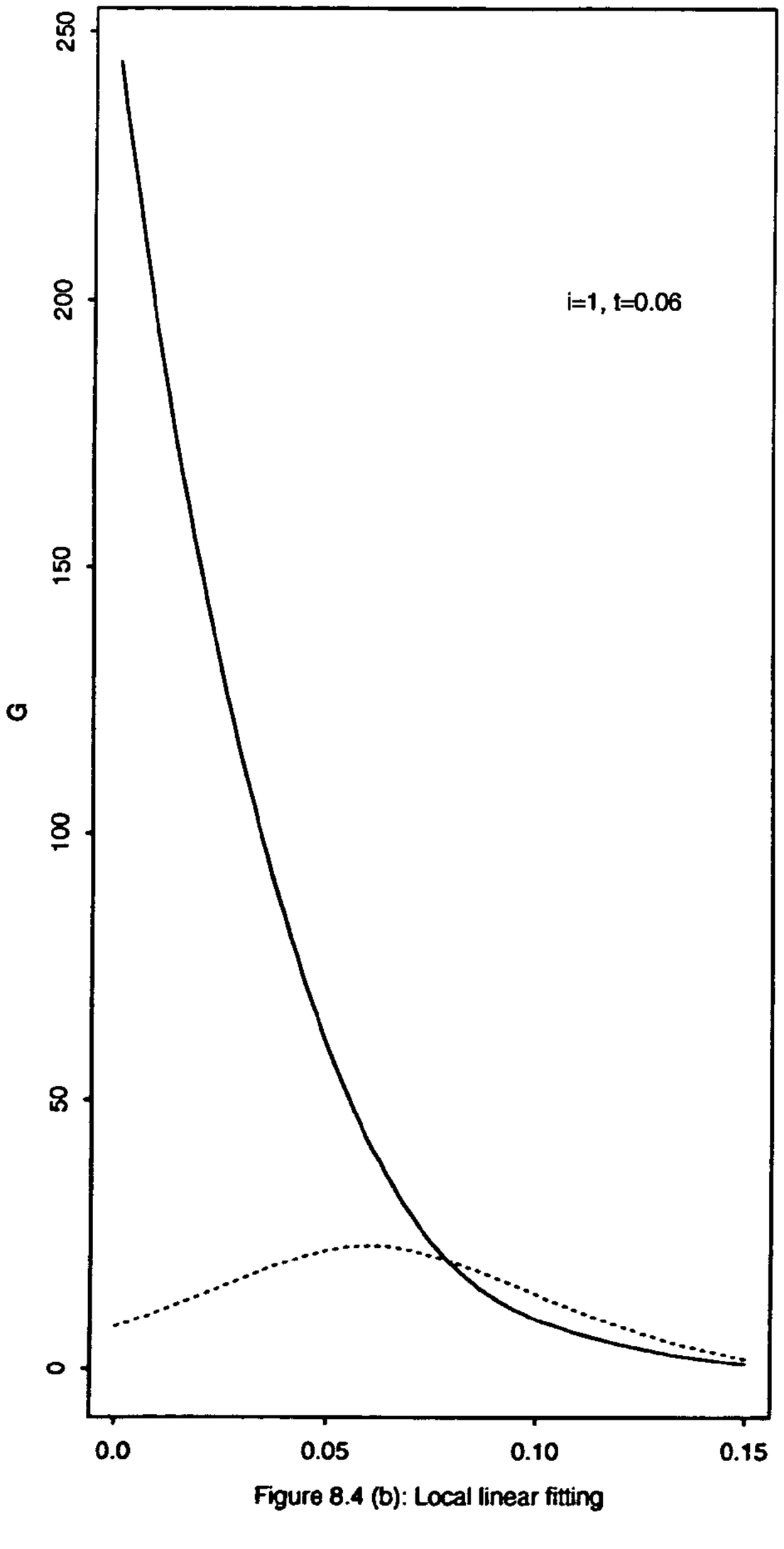


Figure 8.4 (b): Local linear fitting

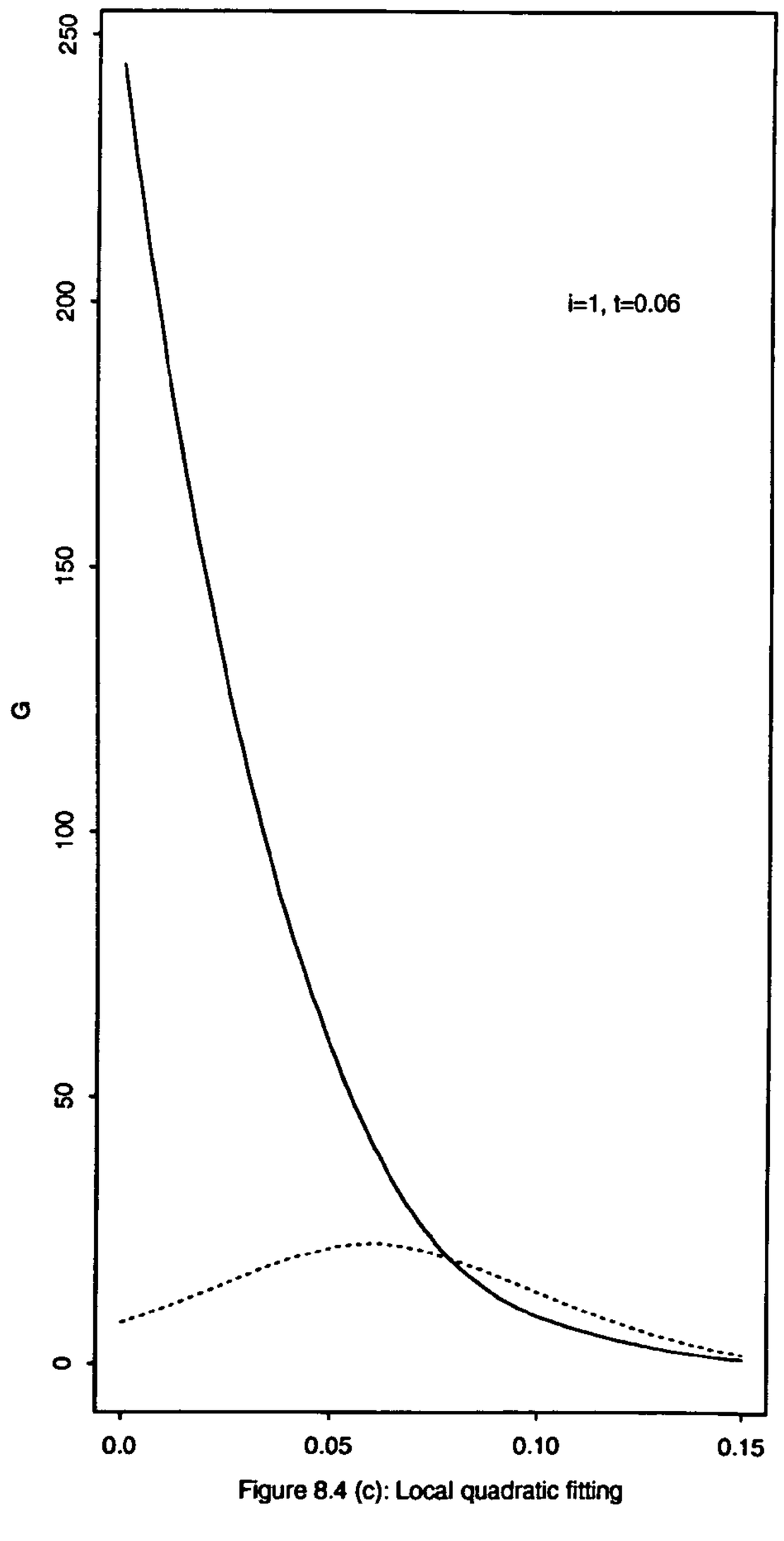


Figure 8.4 (c): Local quadratic fitting

and local quadratic fitting) are not always same, although local linear and local quadratic fittings are remarkably the same by comparing all the (b) and (c) of Figures.

So far, it seems we can make some general points about designing local polynomial spline which brings difference from local constant fitting: selecting penalty term $R(g, g_1, ..., g_m)$ of equation (8.4) with multi-parameter smoothing may give competitive smoothing spline. If specifying the jth derivative of g as g_j in $R(g, g_1, ..., g_m)$ as equation (8.5), the local polynomial fitting makes nothing different from local constant fitting, and the real difference resulting from multi-parameter penalty. However, if regarding all g_j in $R(g, g_1, ..., g_m)$ as independent each other, then local linear fitting at least gives slightly different results in terms of approximate weights and bandwidths from local constant fitting. It is not especially clear, however, whether the differnce is particularly sensible nor whether it would be worthwhile to follow up in practice.

8.6 Integrated Mean Squared Error

To study the closeness property of smoothed curves to the true curve, the widely used measures are pointwise measure MSE and global measure like the MISE. As no boundary modifications are necessary for interior and boundary points, the bias of the equivalent kernel for polynomial fit is expected to be of order $O(h^{2k})$ where the bandwidth h of equivalent kernel is $\left(\frac{\lambda}{a(t)}\right)^{\frac{1}{2k}}$.

Much work should be done for this new spline smoothing in this field, these include

(1) what is the rate that "global" measure converges to zero here? In local constant fit, Rice and Rosenblatt (1983), Eubank(1988) studied smoothing spline and showed that IMSE is dominated by contributions from the boundary points which affects the rate of convergence. However, Nussbaum (1985) pointed out that spline smoothing lead to optimal properties and discussed by Cox (1984).

(2) What is the minimax efficiency of this kind of smoothing spline? In the local constant fit, a minimax type approach to the question of rates of convergence was done by Nussbaum (1985) and state that the conventional smoothing spline is not asymptotically minimax in certain sense whatever the choice of the smoothing parameter λ. On the other hand, Fan(1993) introduce a kernel-type smooth version of local linear regression estimators and addressed its high minimax efficiency.

(3) What is more exact and detailed approach for the weight functions than that of kernel here? In the local constant fit, Messer's (1991) Fourier analysis gives a high order approximation to the weight functions, and Nychka (1995), by using a slightly different approach, showed that the absolute value of the spline weight function decreases exponentially away from its center.

The local linear fit is used as a main representative of local polynomial fit, and it is convenient to use this in discussing asymptotic properties of polynomial fit.

Following the previous model and assumptions, and concentrating on local linear fit with 2nd derivative penalized, write the IMSE, I_n, in terms of its variance and squared bias

$$\begin{aligned} I_n &= E \int_0^1 |f(t) - g(t)|^2 dt \\ &= \int_0^1 |f(t) - Eg(t)|^2 dt + \int_0^1 Var(g(t))\, dt \end{aligned} \tag{8.78}$$

Denoting the Fourier coefficients of true regression function f by

$$a_j = \int_0^1 f(t) exp(-2\pi ijt) dt.$$

The Fourier series of f is

$$f(t) = \sum_{k=-\infty}^{+\infty} a_k exp(2\pi ikt).$$

Under condition A and for $k = 2$ and $m = 1$

$$g(t) \approx \sum_{|j| \leq \frac{n-1}{2}} \frac{\beta_j}{a(t) + \lambda j^4} exp(2\pi ijt). \tag{8.79}$$

First consider the bias component

$$\begin{aligned} B_n^2 &= \int_0^1 |f(t) - Eg(t)|^2 \, dt \\ &= \sum_{|r| > \frac{n-1}{2}} |a_r|^2 + \int_0^1 \sum_{|j| \leq \frac{n-1}{2}} |a_j - \frac{E\beta_j}{a(t) + \lambda j^4}|^2 dt \end{aligned} \tag{8.80}$$

Theorem 8.6.1: If $\lambda \to \infty$ and $n^4\lambda \to \infty$ as $n \to \infty$ under Condition A and $|a_j|^2 = O(|j|^{-5-\delta})$ for some $\delta > 0$,

$$\begin{aligned} B_n^2 &= \int_0^1 |f(t) - Eg(t)|^2 \, dt \\ &= a_0^2 \int_0^1 (1 - \frac{1}{a(t)})^2 dt + \sum_{|j| \leq \frac{n-1}{2}} \int_0^1 |a_j|^2 \frac{\lambda^2 j^8}{(a(t) + \lambda j^4)^2} dt \\ &+ O(n^{-5-\delta}\lambda^{-\frac{1}{4}} + n^{-5-\delta}). \end{aligned} \tag{8.81}$$

if $|a_j|^2 \sim |j|^{-5-\delta}$, $0 < \delta < 4$ then

$$B_n^2 = O(\lambda^{1+\frac{\delta}{4}}) \tag{8.82}$$

if $\sum |a_j|^2 j^8 < \infty$ then

$$B_n^2 = O(\lambda^2). \tag{8.83}$$

Consequently, the bias is $O(\lambda^2)$, viz., $O(h^4)$ in the latter instance.

Proof: since $f \in W_2^2[0,1]$ assume that the a_j decay algebraically, then for some $\delta > 0$ (Chapter 3, Eubank) the first term of (8.80) is

$$\sum_{|j|>\frac{n-1}{2}} |a_j|^2 = O(n^{-(4+\delta)}) \tag{8.84}$$

and the $j = 0$ term in the second sum in B_n^2 is

$$\begin{aligned}
&\int_0^1 |a_0 - \frac{E\beta_0}{a(t)}|^2 dt \\
&\leq a_0^2 \int_0^1 (1 - \frac{1}{a(t)})^2 dt + \int_0^1 \frac{1}{a(t)^2} |a_0 - E\beta_0|^2 dt \\
&\int_0^1 \int_0^1 (1 - \frac{1}{a(t)})^2 dt + |a_0 - E\beta_0|^2 \\
&= \int_0^1 (1 - \frac{1}{a(t)})^2 dt + |\sum_{s\neq 0} a_{sn}|^2 \\
&\leq \int_0^1 (1 - \frac{1}{a(t)})^2 dt + O(n^{-5+\delta}).
\end{aligned} \tag{8.85}$$

The last inequality uses calculations of equations (33) and (38) of Chapter 3 of Eubank (1988).

Consequently, from (8.40), it suffices to the inside part of the integral about t in the second term of (8.80) ($j \neq 0$)

$$\begin{aligned}
&\sum_{1\leq|j|\leq\frac{n-1}{2}} |a_j - E\beta_j \frac{1}{a(t) + \lambda j^4}|^2 \\
&= \sum_{1\leq|j|\leq\frac{n-1}{2}} |a_j - \sum_l a_{j+ln} \frac{1}{a(t) + \lambda j^4}|^2 \\
&= \sum_{1\leq|j|\leq\frac{n-1}{2}} |a_j (1 - \frac{1}{a(t) + \lambda j^4}) \\
&+ \sum_{l\neq 0} a_{j+ln} \frac{1}{a(t) + \lambda j^4}|^2
\end{aligned} \tag{8.86}$$

Use the fact that for any two complex number z_1 and z_2,

$$|z_1 + z_2|^2 = |z_1|^2 + |z_2|^2 + 2Re z_1 \bar{z}_2$$

where Re(z) is the real part of the complex number z. Thus the above sum is

$$\sum_{1\leq|j|\leq\frac{n-1}{2}} |a_j|^2 |1 - \frac{1}{a(t) + \lambda j^4}|^2$$

$$
\begin{aligned}
+ \quad & \sum_{1\le|j|\le\frac{n-1}{2}} |\frac{1}{a(t)+\lambda j^4}|^2 |\sum_{l\neq 0} a_{j+ln}|^2 \\
+ \quad & 2Re[\sum_{1\le|j|\le\frac{n-1}{2}} a_j(1-\frac{1}{a(t)+\lambda j^4}) \\
\cdot \quad & \sum_{l\neq 0} \overline{a_{j+ln}} \frac{1}{a(t)+\lambda j^4}]
\end{aligned} \tag{8.87}
$$

As the last sum is dominated by the squares of the moduli of the other two in (8.87) through the use of the Cauchy-Schwarz inequality, only the first two are considered with integral about t together. On using Cauchy-Schwarz inequality, then

$$
\begin{aligned}
& |\sum_{l\neq 0} a_{j+ln}|^2 \sum_{1\le|j|\le\frac{n-1}{2}} \int_0^1 \Big(\frac{1}{a(t)+\lambda j^4}\Big)^2 dt \\
& \le O(n^{-(5+\delta)}) \sum_{1\le|j|\le\frac{n-1}{2}} \int_0^1 \Big(\frac{1}{a(t)+\lambda j^4}\Big)^2 dt \\
& \le O(n^{-(5+\delta)}) \int_0^1 \int (a(t)+\lambda x^4)^{-2} dx dt \\
& = \quad O(n^{-(5+\delta)})(n\lambda^{1/4})^{-1}
\end{aligned} \tag{8.88}
$$

The last inequality arises from the use of an integral estimate to show that

$$
\sum_{|j|\le(n-1)/2} (a(t)+\lambda j^4)^{-2} \sim a(t)^{-2}(\lambda/a(t))^{-1/4} \int_{-\infty}^{+\infty} (1+y^4)^{-2} dy.
$$

Thus, provided $n\lambda^{\frac{1}{4}} \to \infty$ as $n \to \infty$, the asymptotic properties of B_n^2 will be determined by the first term with integral about t of (8.86). Now consider

$$
\begin{aligned}
& \Sigma_{\ 1\le|j|\le\frac{n-1}{2}} \int_0^1 |a_j|^2 |1-\frac{1}{a(t)+\lambda j^4}|^2 dt \\
= \quad & \sum_{1\le|j|\le\frac{n-1}{2}} \int_0^1 |a_j|^2 |\frac{a(t)-1+\lambda j^4}{a(t)+\lambda j^4}|^2 dt \\
\le \quad & \sum_{1\le|j|\le\frac{n-1}{2}} \int_0^1 |a_j|^2 \frac{\lambda j^4}{a(t)+\lambda j^4}|^2 dt
\end{aligned} \tag{8.89}
$$

Breaking the last sum into two parts the first corresponds to $|j| \le \lambda^{-\frac{1}{4}}$, then

asymptotically, B_n^2 will have the same order as

$$\lambda^2 \sum_{1 \leq |j| \leq \lambda^{\frac{-1}{4}}} |a_j|^2 j^8 + \sum_{|j| > \lambda^{-\frac{1}{4}}} |a_j|^2 \tag{8.90}$$

for $|a_j|^2 \sim |j|^{-(5+\delta)}$ for $0 < \delta < 4$, the first sum in the expression of (8.90) behaves like

$$\lambda^2 \int_0^{\lambda^{-\frac{1}{4}}} x^{-(5+\delta)} \, x^8 \, dx \; \propto \; \lambda^{\frac{4+\delta}{4}}$$

and the second sum as

$$\int_{\lambda^{-\frac{1}{4}}}^{\infty} x^{-(5+\delta)} \, dx \; \propto \; \lambda^{\frac{4+\delta}{4}}$$

So the squared bias B_n^2 is $O(\lambda^{\frac{4+\delta}{4}})$.

Finally, if $\sum_j |a_j|^2 j^8 < \infty$, the first sum of (8.90) is $O(\lambda^2)$ while the second will decay at a rate faster than λ^2, hence the B_n^2 is $O(\lambda^2)$. This establishes the theorem 8.6.1.

Now consider the variance portion of I_n, through use of the fact that

$$Cov(\beta_j, \beta_u) = \begin{cases} \frac{\sigma^2}{n} & \text{if } j = u \\ 0 & \text{otherwise} \end{cases}$$

Combine this with previous approximations to give

$$\begin{aligned} V_n &= \int_0^1 Var(g(t)) \, dt \\ &= \frac{\sigma^2}{n} \sum_{|j| \leq \frac{n-1}{2}} \int_0^1 (a(t) + \lambda j^4)^{-2} dt \end{aligned} \tag{8.91}$$

This, together with Theorem 8.6.1 gives the following theorem

Theorem 8.6.2: Under the conditions of Theorem 8.6.1, if

$$|a_j|^2 \sim |j|^{-5-\delta}, 0 < \delta < 4$$

then

$$I_n \sim C_1\lambda^{1+\frac{\delta}{4}} + C_2(n\lambda^{\frac{1}{4}})^{-1} \tag{8.92}$$

where C_1 and C_2 are positive constants and asymptotically optimal risk is $O(n^{-\frac{4+\delta}{5+\delta}})$.

If

$$\sum |a_j|^2 j^8 \ < \ \infty$$

then

$$I_n \sim C_1\lambda^2 + C_2(n\lambda^{\frac{1}{4}})^{-1} \tag{8.93}$$

and asymptotically optimal risk is $O(n^{-\frac{8}{9}})$.

The latter case arises when $f(t)$ is very smooth in the sense that $f \in W_2^4[0,1]$.

Like local constant fitting, the essential conclusion of Theorem 8.6.2 is that, if $n \to \infty$ and $\lambda \to 0$ in such a way that $n\lambda^{1/4} \to \infty$, $g(t)$ is a consistent estimator of $f(t)$. However, this condition does not depend on λ_1.

Chapter 9

Conclusions and Possible Future Work

With increasing applications in practice and advances made in areas of density estimation and regression mean smoothing, smoothing regression quantiles is increasingly drawing the attention of both mathematical and applied statisticans. Nonparametric approach as compared to parametric is more flexible on the model structure and related distribution assumptions, but not for possible boundary influence and sometimes oversmoothing on some peak. Application of advanced local polynomial kernel regression techniques, and search for new bandwidth selection skills form essential part of research in smoothing regression quantiles. A summary of the progress made in this study towards achieving the good theoretic results and application performance, and some recommendations for future work are given in the following two sections.

9.1 Kernel Smoothing

Nonparametric kernel approach, as in smoothing density function and in smoothing regression mean and quantiles, is theoretically convenient and effective in applications particularly when it is inappropriate to make strong distributional assumptions such as in most of biometrical data. The MSE of kernel estimators are derived, and local linear versions have more concise MSE than local constant versions, when fitting local constant versions is more convenient in computation than fitting local linear versions. It is seen in practice, that both fittings performs equally good at interior points while local linear fittings have better performance in boundary points.

Furthermore, kernel smoothing, except for check function, of regression quantiles, guarantees non-crossing quantiles curves, which is an important phenomenon in favour of this approach. The reason for this is that for (i) double-kernel method has positive second parameter in Y-direction which keeps increasing the direction of tangential motion of quantile function about p, (ii) the semi-parametric method takes advantage of non-crossing quantiles of known parametric distribution, and (iii) two-step method is a simple smoothing of conditionally empirical quantiles.

Naturally, the choice of bandwidth has a central role in kernel smoothing methods.

Unlike smoothing regression mean, the special check function is inconvenient when standard bandwidth techniques are applied particularly in such as cross-validation, as often it is required to smooth several quantiles at one time. Here a rule-of-thumb method is proposed and computational algorithms are established to select bandwidths. The proposed selection rule works well in practice whether one uses a plug-in, single-kernel smoothing or double-kernel smoothing, and that

two-step method is fastest followed by double-kernel method, which are the most effectice two kernel methods in practice. It is found that small changes in the bandwidth have little effects on the percentiles curves.

9.2 Further Work

Several interesting points arise from this study which are left for future work, however, below are some areas where further development and investigation are required.

1. Bandwidth Selection

Study of bandwidths and its estimators require more attention from researchers in the field, particularly estimators of h_1 in plug-in method and h_2 of double-kernel when smoothing conditional distribution and quantiles.

2. Optimal Semi-Parametric Method

Semi-parametric method of Chapter 5 is strongly related to parametric transformation for specific distribution, and a natural approach is to look for optimal transformation based on data.

For a fixed parametric transformation, what is the optimal distribution of the transformed variable Z , i.e. the form of $f_Z(z;x)$ and vice-versa. It is worth mentioning that power transformation with Gamma distribution i.e. $Y^\lambda|x \approx \Gamma(\alpha, 1/\alpha)$ is subject of a study in Bristol. Also, a different parametric transformation $Z = T_\psi(Y)$ with parameter ψ could be used. Assume that Z has distribution $g_Z(z;x)$ with p-quantile φ_p, then $q_p(x;\psi(x)) = T_\psi^{-1}(\varphi_p)$, and an

alternative estimator for $q_p(x)$ may be defined as

$$\bar{q}_p(x) = T^{-1}_{\hat{\psi}(x)}(\varphi_p).$$

3. High-Degree Polynomial and Spline Approaches

The work in this thesis, focussed on local linear fitting extension to higher order polynomial fitting, will be advantageous when smoothing regression quantiles, though order of the polynomial is an additional problem. Certainly the amount of calculation increases and requires developing new computational algorithms.

Alternatively, based on Cox (1988) or Jones (1988), local constant spline approach of Koenker, Portnoy and Ng's (1992) in computing quantiles may be developed to obtain local polynomial spline fitting of quantiles. Simply one may start from local polynomial fitting regression mean as Chapter 8, then extend to the concept to robust spline and quantile spline.

Moreover, for minimization problem based on check function, one may approximate the non-differentiable $\rho(.)$ by a smooth $\rho_\epsilon(.)$ then obtain the solution using Taylor expansion of $\rho_\epsilon(.)$ (scoring method).

4. Wavelet Regression Quantile

Wavelet method of smoothing curve estimation is one of current popular tools for density and regression estimation. It may be developed and extended to regression quantile etimation

$$y_i = f(t_i) + \epsilon_i, \quad i = 1, ..., n,$$

as in Chapter 1 but ϵ_i is independent with zero p-quantile given $0 < p < 1$, so that f is the unknown p-quantile function of interest and t_i may be equispaced

points on the interval. The aim is to recover the f from $\{t_i, y_i\}_1^n$.

Further, wavelet estimators generally capture the local features such as sharp bumps and discontinuities that quantile curves are full of sometimes in practice.

5. Robustness

Comparing to smoothing regression mean estimators, smoothing regression quantiles might be more robust. A simple case is considered here.

Suppose that a large, random batch of mixed "good" and "bad" pairs of independent observations (x_i, y_i) $i = 1, 2, ..., n$ are available for estimating the conditional mean $m(x) = E(Y|X = x)$, and assume that a pair (x_i, y_i) is "bad" with probability ϵ and "good" with probability 1-ϵ , and that (x_i, y_i) are distributed as:

$$(X, Y) \sim \begin{cases} N(0, 0, r, \sigma_1^2, \sigma_2^2) & \text{if } (x_i, y_i) \text{ is good} \\ N(0, 0, r, k\sigma_1^2, k\sigma_2^2) & \text{if } (x_i, y_i) \text{ is bad for } k > 0 \end{cases}$$

In other words, (x_i, y_i) are independent with the common underlying "contaminative distribution":

$$F(x, y) = (1 - \epsilon)F_1(x, y) + \epsilon F_2(x, y)$$

then the conditional distributions $F_1(y|x)$ and $F_2(y|x)$ are $N(r(\sigma_2/\sigma_1)x, \sigma_2^2(1 - r^2))$ and $N(r(\sigma_2/\sigma_1)x, k\sigma_2^2(1 - r^2))$ respectively. And

$$m(x) = q_{1/2}(x) = E(Y|X = x) = r(\sigma_2/\sigma_1)x$$

$$Var(Y|X = x) = \sigma_2^2(1 - r^2)(1 - \epsilon + \epsilon k)$$

$$f(q_{1/2}(x)|x) = (1 - \epsilon)f_1(q_{1/2}(x)|x) + \epsilon f_2(q_{1/2}(x)|x)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1 - r^2}}(1 - \epsilon + \frac{\epsilon}{\sqrt{k}}).$$

Because

$$\{F^{2,0}(q_{1/2}(x)|x)/f(q_{1/2}(x)|x)\}^2 = (q''_{1/2}(x))^2 = 0,$$

the estimate $\hat{m}(x)$ resulting from fitting local linear kernel mean has identical interior bias as the estimate obtained using constant kernel fitting median (Jones and Hall, 1990) or linear kernel fitting median (Fan, Hu and Truong, 1995) where as the variances are respectively proportional to

$$p(1-p)/\{f(q_{1/2}(x)|x)\}^2 \quad and \quad Var(Y|X=x).$$

As $k \to \infty$ the latter tends to infinity, but the former converges to a limit, and in this sense $\hat{q}_{1/2}(x)$ is better than $\hat{m}(x)$ as an estimator of $m(x)$.

6. Application of Local Polynomial Fitting in Survival Analysis

In applications, local polynomial approach may be used with likelihood function in medical models and survival analysis. To illustrate this a simple example of estimating confidence regions of survival function modelled using Cox proportional hazards model is discussed below.

Cox Model in Censored Survival Data: Suppose that the observations are $(\mathbf{x_i}, Y_{ij}, \delta_{ij})$, where $\mathbf{x_i}$ is ith treatment of combination levels of d drugs; T_{ij} and C_{ij} are survival and censored time of the jth experimental unit under treatment combination $\mathbf{x_i}$. Define $Y_{ij} = min(T_{ij}, C_{ij})$; and

$$\delta_{ij} = \begin{cases} 1 & \text{if } T_{ij} \leq C_{ij} \\ 0 & \text{otherwise} \end{cases}$$

for $j = 1$,..., m_i and $i = 1$,..., n.

Let $p(t; g(\mathbf{x}))$ denote the probability density function of the uncensored failure times at the treatment combination $\mathbf{x}$. Assume the hazard functions of the uncensored survival times satisfy the Cox proportional hazards model $h(t; \mathbf{x}) =$

$h(t_0)exp[-g(\mathbf{x})]$, when $\mathbf{x}$ equal to vector ϕ which corresponds to the control group, $g(\phi) = 0$. Denote the survival function corresponding to $p(t; g(\mathbf{x}))$ by S(t; $\mathbf{x}$) and $S_0(t) = S(t; \mathbf{x})_{\mathbf{x}=\phi}$. Then

$$p(t; g(\mathbf{x})) = h_0(t)exp[-g(\mathbf{x})][S_0(t)]^{exp[-g(\mathbf{x})]} \tag{9.1}$$

$$S(t; \mathbf{x}) = [S_0(t)]^{exp[-g(\mathbf{x})]}. \tag{9.2}$$

For known survival function of the control group $S_0(t)$, the goal is to estimate $g(\mathbf{x})$ nonparametrically, here $g(\mathbf{x})$ is the negative-log-relative risk. Actually, estimating the overall surface $g(\mathbf{x})$, the location of the global optimum. For example, the jth treatment combination of levels A and B is given to each of n_i animals, for $i = 1, ..., n$ and n is the total number of treatment combinations. The hazard functions of the uncensored survival times are modeled by Cox proportional hazards model. Estimates of treatment combinations of level A and B that correspond to maximum survival time are required. Since an estimate of the global optimum of $g(\mathbf{x})$ for $\mathbf{x}$ within the experimental region and its confidence regions are ultimate desire, then smooth estimate of derivatives of $g(\mathbf{x})$ are ultimately needed here.

If the censoring mechanism is independent of the failure times and the dose level, then the log-likelihood for the sample (for right-continuous survival functions), up to a constant, is

$$\sum_{ij}\Big(\delta_{ij}log\{p(Y_{ij}; g(\mathbf{x_i}))\} + (1-\delta_{ij})log\{S(Y_{ij}; g(\mathbf{x_i}))\}\Big)$$

$$= \sum_{ij}\Big(\delta_{ij}log\{h_0(Y_{ij})\} - \delta_{ij}g(\mathbf{x_i}) + exp[-g(\mathbf{x_i})]log\{S_0(Y_{ij})\}\Big)$$

The local constant kernel-weight version for estimating $\lambda = g(\mathbf{x})$ maximizes the following sum respect to λ:

$$\sum_{ij} K(\frac{\mathbf{x}-\mathbf{x_i}}{h})\Big(\delta_{ij}log\{h_0(Y_{ij})\} - \delta_{ij}\lambda + exp(-\lambda)log\{S_0(Y_{ij})\}\Big) \tag{9.3}$$

and the local linear kernel-weight version for estimating λ and its derivative maximizes the following sum respect to λ and λ_1:

$$\sum_{ij} K(\frac{\mathbf{x}-\mathbf{x_i}}{h})\Big(\delta_{ij} log\{h_0(Y_{ij})\} - \delta_{ij}(\lambda + \lambda_1(\mathbf{x}-\mathbf{x_i})) + exp(-\lambda + \lambda_1(\mathbf{x}-\mathbf{x_i}))log\{S_0(Y_{ij})\}\Big)$$

Extension to regression quantiles curves for censored data in medicine, or survival analysis is left for future work.

Bibliography

[1] Ahlberg, J.H., Nilson, E.N. and Walsh, J. (1967) *The Theory of Spline and Their Applications.* New York: Academic Press.

[2] Altman, N.S. (1990) "Kernel Smoothing of Data with Correlated Errors", *Journal of the American Statistical Association*, 85, 749-759.

[3] Azzalini, A. (1981) "A Note on the Estimation of a Distribution Function and Quantiles by a Kernel Method", *Biometrika*, 68, 326-32.

[4] Bahadur, R.R. (1966) "A Note on Quantiles in Large Samples". *Annals of Mathematical Statistics*, 37, 577-580.

[5] Bhattacharya, P.K. & Gangopadhyay, A.K. (1990) "Kernel and Nearest-Neighbor Estimation of a Conditional Quantile." *The Annals of Statistics*, 18, 1400-1415.

[6] Bowman, A.W (1985) "A Comparative Study of Some Kernel-Based Non-parametric Density Estimators", *Journal of Statistical Computation and Simulation*, 21, 313-327.

[7] Bracewell, R.(1965) *The Fourier Transform and its Application*, McGraw-Hill Book Company.

[8] Breiman, L. & Friedman, H. (1985) "Estimating Optimal Transformations for Multiple Regression and Correlation", with Comments, *Journal of the American Statistical Association*, 80, 580-619.

[9] Chaudhuri, P. (1991) "Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation", *Annals of Statistics*, 2, 760–777.

[10] Chambers, J.M.and Hastie, T.J.(1992) *Statistical Models in S.* Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, California.

[11] Cheng, K.F. (1983), "Nonparametric Estimators for Percentiles Regression Functions", *Communications in Statistics. Theory and Methods*, 12 681-692.

[12] Cheng, K.F. (1984) "Nonparametric Estimation of Regression Function Using Linear Combinations of Sample Quantile Regression Function", *Sankhyā*, Ser .A 46, 287-302.

[13] Chu, C.-K. and Marron, J.S. (1992) "Choosing a Kernel Regression Estimator(with Discussion and Rejoinder)", *Statistical Science*, 6, 404-436.

[14] Cleveland, W. (1979) "Robust Locally Weighted Regression and Smoothing Scatterplots". *Journal of the American Statistical Association*, 74, 829–836.

[15] Cleveland, W., and Devlin, S.J. (1988) "Locally-Weighted Regression and Smoothing Scatterplots", *Journal of the American Statistical Association*, 83, 597–610.

[16] Cleveland, W., and Loader, C. (1995) "Smoothing by Local Regression: Principles and Methods". *Computational Statistics*, to appear.

[17] Cole, T.J. (1988) "Fitting Smoothed Centile Curves to Reference Data", *Journal of the Royal Statistical Society*, Series A, 151, 385–418.

[18] Cole, T.J., and Green, P.J. (1992) "Smoothing Reference Centile Curves: the LMS Method and Penalized Likelihood", *Statistics in Medicine*, 11, 1305–1319.

[19] Cox, D. (1984) "Multivariate Smoothing Spline Functions". *SIAM Journal on Numerical Analysis*, 21, 789-813.

[20] Cox, D. R. (1988) In the Discussion to Cole (1988), *Journal of the Royal Statistical Society*, Series A, 151, 411.

[21] Craven, P.and Wahba, G.(1979) "Smoothing Noisy Data with Spline Functions". *Numerische Mathematik*, 31, 377-390.

[22] Crowley, J. and Hu, M. (1977) "Covariance Analysis of Heart Transplant Data", *Journal of the American Statistical Association*, 72, 27-36.

[23] De Boor, C.(1978) *A Practical Guide to Splines.* New York: Springer-Verlag.

[24] Dold, E. and Eckmann, B. (1979) *Smoothing Techniques for Curve Estimation.* Lecture Notes in Math., Springer-Verlag.

[25] Efron, B. (1991) "Regression Percentiles Using Asymmetric Squared Error Loss", *Statistica Sinica*, 1, 93-125.

[26] Ellis, T.M.R. (1990) *Fortran 77 Programming, with an introduction to the Fortran 90 Standard.* Second Edition, Addison-Wesley Publishing Company.

[27] Eubank, R.L.(1988) *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker.

[28] Fan, J. (1992) "Design-Adaptive Nonparametric Regression", *Journal of the American Statistical Association*, 87, 998–1004.

[29] Fan, J. (1993) "Local Linear Regression Smoothing and Their Minimax Efficiencies", *The Annals of Statistics*, 21, 196–216.

[30] Fan, J., and Gijbels, I. (1992) "Variable Bandwidth and Local Linear Regression Smoothers", *The Annals of Statistics*, 20, 2008–2036.

[31] Fan, J., and Gijbels, I. (1995) "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation", *Journal of the Royal Statistical Society*, Series B, 57, 371–394.

[32] Fan, J. Heckman, N.E. & Wand, M.P. (1995) "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood functions", *Journal of the American Statistical Association*, 90 141-150.

[33] Fan, J., Hu, T.-C., and Truong, Y.K. (1994) "Robust Nonparametric Function Estimation", *Scandinavian Journal of Statistics*, 21, 433-446.

[34] Fan, J., Yao, Q., and Tong, H. (1996) "Estimating Measures of Sensitivity of Initial Values to Nonlinear Stochastic System with Chaos", *Biometrika*, 83, 189-206.

[35] Gasser, T. and Muller, H-G. (1984) "Estimating Regression Functions and Their Derivatives by the Kernel Method." *Scandinavian Journal of Statistics* 11, 171-185.

[36] Goldstein, H. and Pan, H. 91992) "Percentile Smoothing Using Piecewise Polynomials, with Covariates". *Biometrics*, 48, 1057-1068.

[37] Good, I.J. and Gaskins, R.A. (1971) "Nonparametric Roughness Penalties for Probability Densities". *Biometrika*, 58, 255-277.

[38] Green, P.J. (1984) "Iterated Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alteratives (with Discussion)," *Journal of the Royal Statistical Society,* B, 46, 149-192.

[39] Green, P.J (1988) In the Discussion to Cole (1988) , *Journal of the Royal Statistical Society*, Series A, 151, 410.

[40] Green, P.J. and Silverman, B.W. (1994) *Nonparametric Regression and Genalized Linear Models, A Roughness Penality Approach.* London: Chapman and Hall.

[41] Härdle, W. (1990) *Applied Nonparametric Regression*, Cambridge: Cambridge University Press.

[42] Hart, J.D. & Wehrly, T.E. (1986), "Kernel Regression Estimation Using Repeated Measurements Data," *Journal of the American Statistical Association*, 81, 1080-1088.

[43] Hastie, T., and Loader, C. (1993) "Local Regression: Automatic Kernel Carpentry", *Statistical Science*, 8, 120–143.

[44] Hastie, T., and Tibshirani, R. (1990) *Generalized Additive Models.* London: Chapman and Hall.

[45] Healy, M.J.R., Rasbash, J. and Yang, M. (1988). "Distribution-Free Estimation of Age-Related Centiles". *Annals of Human Biology*, 15, 17-22.

[46] Hendricks, W. and Koenker, R. (1992) "Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity", *Journal of the American Statistical Association*, 58-68.

[47] Herrmann, E. Gasser, T. & Kneip, A. (1992) "Choice of Bandwidth for Kernel Regression When Residuals are Correlated", *Biometrika*, 79, 783-795.

[48] Hill, P.D. (1985) "Kernel estimation of a Distribution function", *Communications in Statistics. Theory & Methods*, 14, 605-620 .

[49] Hogg, R. V. (1975) "Estimating of Percentile Regression Lines Using Salary Data", *Journal of the American Statistical Association*, 70, 56-59.

[50] Horavth,L. (1988) "Asymptotics of Conditional Empirical Processes", *Journal of Multivariate Analysis*, 26, 184-206.

[51] Huber, P.J. (1981) *Robust Statistics*. John Wiely and Sons, Inc.

[52] Hunsberger, S. (1994) "Semiparametric Regression in Likelihood-Based Models", *Journal of the American Statistical Association*, 89, 1354-1365.

[53] Isaacs, D., Altman, D.G., Tidmarsh, C.E., Valman, H.B., and Webster, A.D.B. (1983) "Serum Immunoglobin Concentrations in Preschool Children Measured by Laser Nephelometry: Reference Ranges for IgG, IgA, IgM", *Journal of Clinical Pathology*, 36, 1193–1196.

[54] Janssen, P. and Veraverbeke, N. (1987) "On Nonparametric Regression Estimators Based on Regression Quantiles", *Communication in Statistics. Theory and Methods*, 16 383-396.

[55] Jones, M.C. (1988) In the discussion to Cole (1988), *Journal of the Royal Statistical Society*, Series A, 151, 412-413.

[56] Jones, M.C. and Hall, P. (1990) "Mean Square Error Properties of Kernel Estimates of Regression Quantiles", *Statistics and Probability Letters*, 10, 283-289.

[57] Jones, M.C., Marron, J.S., and Sheather, S.J. (1996) "A Brief Survey of Bandwidth Selection for Density Estimation", *Journal of the American Statistical Association*, 91, 401-407.

[58] Kincaid, D. and Cheny, W. (1990) *Numerical analysis, Mathematics of Scientific Computing.* Brooks/Cole Publishing Company, Pacific Grove, California.

[59] Koenker, R. and Bassett, G.S. (1978) "Regression Quantiles", *Econometrica*, 46, 33–50.

[60] Koenker, R. and Bassett, G.S. (1982) "Robust Tests for Heteroscedasticity Based on Regression Quantiles", *Econometrica*, 50, 43-60.

[61] Koenker, R., Portnoy, S., and Ng, P. (1992) "Nonparametric Estimation of Conditional Quantile Functions", in L_1*-Statistical Analysis and Related Methods* (ed: Y. Dodge) Amsterdam: Elsevier, 217-229.

[62] Lejeune, M.G., and Sarda, P. (1988) "Quantile Regression: A Nonparametric Approach", *Computational Statistics & Data Analysis*, 6, 229–239.

[63] Magee, L., Burbidge, J. B. and Robb, A. L. (1991) "Computing Kernel-Smoothed Conditional Quantiles from Many Observations", *Journal of the American Statistical Association*, 86, 673-677.

[64] Marron, J.S. and Wand, M.P. (1992) "Exact Mean Integrated Squared Error", *The Annals of Statistics*, 20, 712-73 .

[65] Masry, E. & Fan, J. (1993) "Local Polynomial Estiamtion of Regression Functions", Technical Report,No. 2311, Chapel Hill, North Carolina.

[66] Messer, K.(1991) "A Comparison of a Spline Estimate to its Equivalent Kernel Estimate." *The Annals of Statistics*, 2, 817-829.

[67] Mohamed, A.A. Moussa & Mohamed Y. Cheema (1992) "Non-Parametric Regression Regression in Curve Fitting", *The Statistican*, 41, 209-225.

[68] Müller, H.G. (1988) *Nonparametric Analysis of Longitudinal Data*, Berlin, Springer-Verlag.

[69] Mood, A.M. (1950) *Introduction to the Theory of Statistics*, New York: McGraw-Hill Book Co..

[70] Müller, H.G. and Stadtmüller, U. (1987) "Variable Bandwidth Kernel Estimators of Regression Curves", *The Annals of Statistics*, 15, 182-201.

[71] Nussbaum, M. (1985) "Spline Smoothing in Regression Models and Asymptotic Efficiency in L_2". *The Annals of Statistics*, 13, 984-997.

[72] Nychka, D. (1995) "Splines as Local Smoothers". *The Annals of Statistics*, 23, 1175-1197.

[73] Oberhettinger, F. (1990) *Tables of Fourier Transforms and Fourier Transforms of Distributions*, Springer-Verlag.

[74] Pan, H., Goldstein, H. and Yang, Q. (1990) "Nonparametric Estimation of Age-related Centiles Over Wide Age Ranges", *Annals of Human Biology*, 17, 475-81.

[75] Portnoy, S. and Welsh, A.H. (1992) "Exactly What is Being Modelled by the Systematic Component in a Heteroscedastic Linear Regression", *Statistics & Probability Lettter*, 13, 253-258.

[76] Portnoy, S. (1991) "A Regression Quantile Based Statistics for Testing Non-stationarity of Errors, to appear in: *Nonparametric Statistics and Related Topics* (ed: A.K.Md.E.Saleh), North Holland: New York.

[77] Rasbash, J. and Pan, H. (1990) "GROSTAT II: A Program for Estimating Age-Related Centiles", London: World Health Organisation Centre for Growth and Development, University of London.

[78] Reinsch, C. (1967) "Smoothing Spline Function". *Numerische Mathematik,* 10, 177-183.

[79] Rice, J. and Rosenblatt, M. (1981) "Integrated mean squared error of a smoothing spline". *Journal of Approximation Theorys*, 33, 353-369.

[80] Rice, J.and Rosenblatt, M. (1983) "Smoothing Splines: Regression, Derivatives and Deconvolution". *The Annals of Statistics*, 11, 141-156.

[81] Rossiter, J.E. (1991) "Calculating Centile Curves Using Kernel Density Estimation Methods with Application to Infant Kidney Lengths", *Statistics in Medicine*, 10, 1693-1701.

[82] Royston, P. (1991) "Constructing Time-Specific Reference Ranges", *Statistics in Medicine*, 10, 691-695.

[83] Royston, P. and Matthews, J.N.S. (1991) "Estimation of Reference Ranges from Normal Samples", *Statistics in Medicine*, 10, 691-695.

[84] Royston, P., and Altman, D.G. (1994) "Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling" (with discussion), *Applied Statistics*, 43, 429–467.

[85] Ruppert, D., Sheather, S.J., and Wand, M.P. (1995) "An Effective Bandwidth Selector for Local Least Squares Regression", *Journal of the American Statistical Association*, 90, 1257-1270.

[86] Ruppert, D. and Wand, M.P. (1994) "Multivariate Locally Weighted Least Squares Regression", *The Annals of Statistics*, 23, 1346–1370.

[87] Ruppert, D., Wand, M.P. Holst, U. and Hössjer, O. (1995) "Local Polynomial Variance Function Estimation", Technical Report, University of New South Wales.

[88] Sakia, R.M. (1992) "The Box-Cox Transformation Technique: A Review", *The Statistician*, 41, 169-178.

[89] Samanta, M. (1989) "Non-Parametric Estimation of Conditional Quantiles", *Statistics and Probability Letters*, 7, 407–412.

[90] Scallan, A.J. (1992) "Maximum Likelihood Estimation for a Normal/Laplace Mixture Distribution", *The Statistician*, 41, 227-231.

[91] Schucany, W.R. (1995) "Adaptive Bandwidth Choice For Kernel Regression", *Journal of the American Statistical Association*, 90, 535-540.

[92] Silverman, B.W. (1984) "Spline Smoothing: the Equivalent Variable Kernel Method." *The Annals of Statistics* 12, 898-916.

[93] Silverman, B.W. (1985) "Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting (with Discussion)". *Journal of the Royal Statistical Society,* B, 47, 1-52.

[94] Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

[95] Silverman, B.W.(1992) "Should We Use Kernel Methods At All?" *Statistical Science,* 6, 430-433.

[96] Staniswalis, J.G. (1989) "The Kernel Estimate of a Regression Function in Likehood-Based Models", *Journal of the American Statistical Association,* 84, 276-288.

[97] Stone, C.J. (1977) "Consistent Nonparametric Regression", *The Annals of Statistics*, 4, 595–645.

[98] Stute, W. (1986) "Conditional Empirical Processes", *The Annals of Statistics* 14, 638-647.

[99] Thompson, J.R. and Tapia, R. (1990) *Nonparametric Function Estimation, Modeling, and Simulation.* SIAM, Philadelphia.

[100] Thompson, M.L. and Theron, G.B. (1990) "Maximum Likelihood Estimation of Reference Centiles", *Statistics in Medicine*, 9, 539-548.

[101] Tibshirani, R. and Hastie, T. (1987) "Local Likelihood Estimation", *Journal of the American Statistical Association*, 82, 559-568.

[102] Wahba, G. (1975) "Smoothing Noisy Data with Spline Functions. *Numerical Mathematics*, 24, 383-393.

[103] Wahba, G. (1990) *Spline Models for Observational Data.* SIAM, Philadelphia, Pennsylvania.

[104] Van't Hof, M.A., Wit, J.M. and Roede, M.J.(1985) "A Method to Construct Age References for Skewed Skinfold Data, Using Box-Cox Transformation to Normality", *Human Biology*, 57, 131-139.

[105] Wand, M.P. and Jones, M.C. (1993) "Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation". *Journal of the American Statistical Association*, 88, 520-527 .

[106] Wand, M.P., and Jones, M.C. (1995) *Kernel Smoothing*, London: Chapman and Hall.

[107] Wand, M.R., Marron, J.S. & Ruppert, D. (1991) "Transformations in Density Estimation", *Journal of the American Statistical Association*, 86 343-361.

[108] Wang, N. & Ruppert, D. (1995) "Nonparametric Estimation of the Transformation in the Transform-Both-Sides Regression Model", *Journal of the American Statistical Association*, 90, 522-534.

[109] West, M. (1993) "Approximating Posterior Distributions by Mixtures", *Journal of the Royal Statistical Society*, B, 55, 409-422.

[110] Wright, E. (1995) "Quantile Regression for Survival Analysis", Ph.D thesis, University of Glasgow.

[111] Yu, K. and Jones, M.C. (1997a) "Local Linear Regression Quantile Estimation". To appear on *Journal of the American Statistical Association.*

[112] Yu, K. and Jones, M.C. (1997b) "A Comparison of Local Constant and Local Linear Regression Estimation". To appear on Journal of Computational Statistics and Data Analysis.

[113] Yu, K. (1997) "Percent Smoothing by Combining $k-NN$ with Local Linear Fitting". To submit.

[114] Yu, K. and Jones, M.C. (1994) "Local Polynomial Fitting with k Derivative Penalty". Technical report, The Open University.